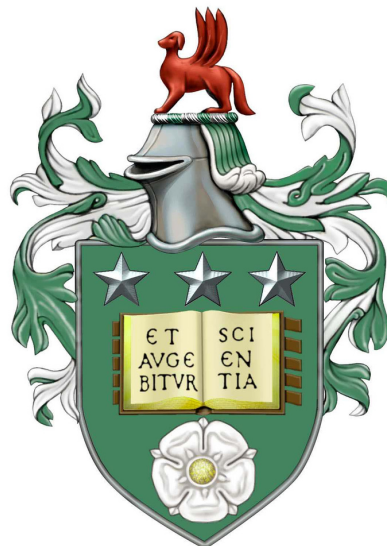


Artificial Intelligence and Patient-Reported Outcome Measures for Predicting Hospital Utilisation and Health-Related Quality of Life During Chemotherapy

Zuzanna Wójcik

A thesis presented for the degree of
Doctor of Philosophy



School of Computer Science
University of Leeds
United Kingdom
September 2025

Intellectual Property and Publication Statements

I confirm that the work submitted is my own, except where work which has formed part of jointly authored publications has been included. My contribution and the other authors to this work has been explicitly indicated below. I confirm that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 2 was based on the following publication:

[1] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2025. Using artificial intelligence to predict patient outcomes from patient-reported outcome measures: a scoping review. *Health and Quality of Life Outcomes*, 23(1), p.37.

I was a primary author of this publication. My contributions were conception, design, data collection (literature search), data analysis, data interpretation, manuscript writing and tables and figures generation. The other authors are PhD supervisors of the project and provided supervisory guidance and editing.

Chapter 4 was based on the following publication:

[2] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2024. Using machine learning to predict unplanned hospital utilization and chemotherapy management from patient-reported outcome measures. *JCO Clinical Cancer Informatics*, 8, p.e2300264.

I was a primary author of this publication. My contributions were conception, design, data analysis, data interpretation, manuscript writing and tables and figures generation. The other authors are PhD supervisors of the project and provided supervisory guidance and editing. The data used in this paper were collected in a clinical trial conducted by Kate Absolom, Lorraine Warrington and Galina Velikova, which is acknowledged and referenced in the publication.

Chapter 5 was based on the following publication:

[3] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2024. Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes During Chemotherapy. In *International Conference on Artificial Intelligence in Medicine* (pp. 101-116). Cham: Springer

Nature Switzerland.

I was a primary author of this publication. My contributions were conception, design, data analysis, data interpretation, manuscript writing and tables and figures generation. The other authors are PhD supervisors of the project and provided supervisory guidance and editing. The data used in this paper were collected in a clinical trial conducted by Kate Absolom, Lorraine Warrington and Galina Velikova, which is acknowledged and referenced in the publication.

Chapter 6 was based on the following publication:

[4] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G., Absolom, K. and Relton, S.D., 2025. Machine Learning Models Predicting Hospital Admissions During Chemotherapy Utilising Longitudinal Symptom Severity Reports and Patient-Reported Outcome Measures. In *Intelligent Health Systems—From Technology to Data and Knowledge* (Vol. 327, pp. 178-182). IOS Press.

I was a primary author of this publication. My contributions were conception, design, data analysis, data interpretation, manuscript writing and tables and figures generation. The other authors provided supervisory guidance and editing as a supervisory team (Vania Dimitrova, Kate Absolom, Lorraine Warrington and Galina Velikova) or as an advisor (Samuel Relton). The data used in this paper were collected in a clinical trial conducted by Kate Absolom, Lorraine Warrington and Galina Velikova, which is acknowledged and referenced in the publication.

Chapter 7 was based on the following publication:

[5] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G., Absolom, K. and Relton, S.D., 2025. Time-Series Patient-Reported Data and LSTM Predicting Hospital Utilisation During Chemotherapy. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)* (pp. 228-234). IEEE.

I was a primary author of this publication. My contribution to this paper was conception, design, data analysis, data interpretation, manuscript writing and tables and figures generation. The other authors provided supervisory guidance and editing as a supervisory team (Vania Dimitrova, Kate Absolom, Lorraine Warrington and Galina Velikova) or as an advisor (Samuel Relton). The data used in this paper were collected in a clinical trial conducted by Kate Absolom, Lorraine Warrington and Galina Velikova, which is acknowledged and referenced in the publication.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

List of Publications

Journal papers:

- [1] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2025. Using artificial intelligence to predict patient outcomes from patient-reported outcome measures: a scoping review. *Health and Quality of Life Outcomes*, 23(1), p.37.
- [2] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2024. Using machine learning to predict unplanned hospital utilization and chemotherapy management from patient-reported outcome measures. *JCO Clinical Cancer Informatics*, 8, p.e2300264.

Conference papers:

- [3] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G. and Absolom, K., 2024. Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes During Chemotherapy. In *International Conference on Artificial Intelligence in Medicine* (pp. 101-116). Cham: Springer Nature Switzerland.
- [4] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G., Absolom, K. and Relton, S.D., 2025. Machine Learning Models Predicting Hospital Admissions During Chemotherapy Utilising Longitudinal Symptom Severity Reports and Patient-Reported Outcome Measures. In *Intelligent Health Systems—From Technology to Data and Knowledge* (Vol. 327, pp. 178-182). IOS Press.
- [5] **Wójcik, Z.**, Dimitrova, V., Warrington, L., Velikova, G., Absolom, K. and Relton, S.D., 2025. Time-Series Patient-Reported Data and LSTM Predicting Hospital Utilisation During Chemotherapy. In *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)* (pp. 228-234). IEEE.

All publications are attached in the Appendix (8.1, 8.2, 8.3, 8.4, and 8.5).

Acknowledgements

This thesis would not have been possible without an endless support of many people, who have made the past four years enjoyable.

Firstly, I would like to express my deepest gratitude to my supervisors: Professor Vania Dimitrova, Professor Galina Velikova, Dr Kate Absolom and Dr Lorraine Warrington. Their expertise, guidance and kind approach provided me with opportunities to develop both as a person and as a researcher. I greatly appreciate the time and effort they have put to support me throughout the past four years.

This research is supported in part by UK Research and Innovation (UKRI) and I am grateful for the funding which allowed me to conduct the research reported in this thesis. Being a part of the UKRI Centre for Doctoral Training (CDT) in Artificial Intelligence for Medical Diagnosis and Care at the University of Leeds has provided me with an amazing research community. I would like to thank the manager of the CDT, the centre directors, the patient representatives and PhD students for the support and motivation. I am also grateful to Dr Samuel Relton for his valuable advice on a part of this research.

I would like to thank patients and clinicians from the eRAPID clinical trial for providing the data for this research. I am also grateful to the clinicians, researchers and patient representatives from the Patient Centred Outcomes Research group from the Leeds Institute of Medical Research at St James's and the Use My Data independent movement of patients, relatives and carers who took their time to discuss this research. Their involvement has helped shaping the direction of this thesis.

I am very fortunate to have such an amazing family and I am grateful for all the help, encouragement and understanding I have received from my parents. I appreciate that you have always been there for me and believed in me unconditionally. I also would like to thank my brother and my grandparents for being very supportive over the past four years.

Harry, I met you half-way through my PhD and you immediately became my biggest cheerleader. I am grateful for your endless love and support. Thank you for motivating me throughout this time and always celebrating my achievements. I cannot wait for our next phase.

Finally, I would like to express appreciation to my best friend Justyna. Thank you for always being there to listen and support. Your drive to achieve your goals is inspiring and has motivated me throughout this PhD.

Abstract

Patient-reported outcome measures (PROMs) are any reports provided by the patient without interpretation from a clinician or anybody else. The use of PROMs in clinical practice can facilitate shared decision-making and self-management of symptoms. PROMs are increasingly being used in Artificial Intelligence (AI) models to predict patient outcomes. However, the lack of guidance on using PROMs in AI models leads to inconsistencies in data pre-processing, model development and model evaluation. This hinders the practical adoption of PROMs and AI models and limits their potential to improve patient outcomes.

Through a literature scoping review, the following research gaps were found: 1) Unclear predictive value of PROMs; 2) Inconsistent data pre-processing methods; 3) Lack of patient and clinician involvement in the study design; 4) Unexplored potentials of time-series PROMs and deep learning (DL) models. To address these gaps, this thesis developed a novel, patient-centric framework for applying AI methods on PROMs. Data collected in eRAPID clinical trial were used to assess machine learning (ML) models and long-short term memory model (LSTM) predicting hospital utilisation and health-related quality of life (HRQoL) changes during chemotherapy. The performance of models including and excluding PROMs and developed on data affected by different pre-processing methods were compared. ML models and LSTM built on longitudinal PROMs and symptom severity reports collected weekly were also explored. Consultations with a clinical oncologist and patient representatives were an essential part of the study design.

The results indicate that PROMs can improve the performance of AI models, which encourages the collection and use of patient-reported data in AI research. Using oversampling techniques can affect the performance of models predicting the same outcomes. This thesis recommends approaches which mitigate the risk of bias (including preventing test set from the effects of data pre-processing) and nested cross-validation pipeline for model development and evaluation. HRQoL outcomes were chosen by patients and a clinical oncologist, which emphasises the importance of stakeholder involvement. Finally, the results indicate that LSTM model can detect sudden changes in data to successfully predict hospital utilisation events within 14 days from a completed symptom severity report, but does not outperform simple ML models for long-term predictions.

Based on these results, a patient-centric framework to develop and evaluate clinically relevant AI models applied on static and time-series PROMs is proposed. This framework can help mitigate bias and support the consistent use of valuable PROMs data in AI research predicting patient outcomes.

Contents

Intellectual Property and Publication Statements	i
List of Publications	iii
Acknowledgements	iv
Abstract	v
List of Figures	xii
List of Tables	xiv
List of Acronyms	xv
Glossary	xvii
1 Introduction	1
1.1 Artificial Intelligence (AI)	1
1.2 Artificial Intelligence in Medicine	1
1.3 Patient-Centred Care	2
1.4 Patient-Reported Outcome Measures (PROMs)	3
1.4.1 History of PROMs	3
1.4.2 Benefits of PROMs	4
1.5 PROMs in AI Research	5
1.6 Aims and objectives	6
1.7 Thesis structure	7
2 Literature Review	8
2.1 Introduction	8
2.2 Methods	9
2.2.1 Search Strategy	9
2.2.2 Inclusion and Exclusion Criteria	9
2.2.3 Review Process	10
2.2.4 Data Extraction and Analysis	10
2.3 Results	12
2.3.1 Study Characteristics	12
2.3.2 Data Pre-Processing	15
2.3.3 Model Development	15
2.3.4 Model Evaluation	16
2.3.5 Clinical Relevance and Adoption	18

2.4	Discussion	18
2.4.1	Study Characteristics	18
2.4.2	Data Pre-Processing	19
2.4.3	Model Development	19
2.4.4	Model Evaluation	19
2.4.5	Clinical Relevance and Adoption	20
2.4.6	Recent Literature	20
2.4.7	Strengths and Limitations	21
2.4.8	Conclusions	22
3	Research Methodology	23
3.1	Introduction	23
3.2	Research Questions	23
3.3	Clinical Context	25
3.3.1	Chemotherapy Toxicity	25
3.3.2	eRAPID Clinical Trial	25
3.4	Dataset	28
3.4.1	Predictor variables	29
3.4.2	Target Variables	29
3.4.3	Data Management	29
3.5	Methodology Overview	32
3.5.1	Description of Research Phases	32
3.5.2	Summary	39
4	Using Baseline PROMs To Predict Hospital Utilisation	40
4.1	Introduction	40
4.2	Methods	42
4.2.1	Methodology for Predicting Hospital Utilisation from Baseline PROMs	42
4.2.2	Dataset	42
4.2.3	Data Pre-processing	44
4.2.4	Model Development	46
4.2.5	Model Evaluation	49
4.2.6	Stakeholder Engagement	49
4.3	Results	50
4.3.1	Hospital Admissions	50
4.3.2	Triage Events	51
4.3.3	Changes to Chemotherapy	53
4.3.4	Significance Testing	56
4.3.5	Variable Importance	56
4.4	Discussion	58
4.4.1	Overall Results	58
4.4.2	Strengths and Limitations	60
4.4.3	Conclusions	61
5	Using Baseline PROMs To Predict HRQoL Changes	63
5.1	Introduction	63
5.2	Methods	64

5.2.1	Methodology for Predicting Health-Related Quality of Life Changes from Baseline PROMs	64
5.2.2	Dataset	65
5.2.3	Data Pre-Processing	67
5.2.4	Model Development	67
5.2.5	Model Evaluation	69
5.2.6	Stakeholder Engagement	69
5.3	Results	70
5.3.1	Physical Well-Being	70
5.3.2	Social Functioning	70
5.3.3	Role Functioning	70
5.3.4	Usual Activities	71
5.3.5	Mobility	71
5.3.6	Significance Testing	71
5.3.7	Variable Importance	74
5.4	Discussion	74
5.4.1	Overall Results	74
5.4.2	Strengths and Limitations	76
5.4.3	Conclusions	76
6	Using ML and time-series PROMs To Predict Hospital Utilisation and HRQoL Changes	78
6.1	Introduction	78
6.2	Methods	80
6.2.1	Methodology for Using Machine Learning to Predict Outcomes from Longitudinal PROMs	80
6.2.2	Dataset	81
6.2.3	Data Pre-processing	82
6.2.4	Model Development	84
6.2.5	Model Evaluation	84
6.2.6	Stakeholder Engagement	85
6.3	Results	85
6.3.1	Hospital Utilisation	86
6.3.2	HRQoL Changes	88
6.3.3	Significance Testing	89
6.3.4	Variable Importance	90
6.4	Discussion	90
6.4.1	Overall Results	90
6.4.2	Strengths and Limitations	92
6.4.3	Conclusions	93
7	Using LSTM and time-series PROMs To Predict Hospital Utilisation and HRQoL Changes	95
7.1	Introduction	95
7.2	Methods	97
7.2.1	Methodology for Using LSTM to Predict Outcomes from Time-Series PROMs	97
7.2.2	Dataset	97
7.2.3	Data Pre-Processing	97

7.2.4	Model Development	99
7.2.5	Model Evaluation	100
7.2.6	Stakeholder Engagement	101
7.3	Results	101
7.3.1	Hospital Utilisation	101
7.3.2	HRQoL Changes	102
7.3.3	Significance Testing	102
7.3.4	Variable Importance	103
7.4	Discussion	106
7.4.1	Overall Results	106
7.4.2	Strengths and Limitations	108
7.4.3	Conclusions	109
8	Discussion	111
8.1	Introduction	111
8.2	Synopsis	111
8.2.1	Summary of Chapters	111
8.2.2	Key Findings	117
8.3	Recommended Framework	118
8.4	Original Contributions	119
8.4.1	Generic Contributions	119
8.4.2	Contributions Related to the Specific Clinical Context for this Research	122
8.5	Strengths	123
8.6	Limitations and Future Work	124
8.7	Reflections	125
8.8	The Future of AI and PROMS	126
	Bibliography	129
	Appendix	165
8.1	Paper 1	165
8.2	Paper 2	181
8.3	Paper 3	200
8.4	Paper 4	206
8.5	Paper 5	225

List of Figures

2.1	PRISMA flow diagram.	12
2.2	Year of publication of all 94 studies (top figure) and distribution of the studies based on their healthcare domain (bottom figure).	13
2.3	Number of papers categorised based on sample sizes and healthcare domain.	14
2.4	Reporting of pre-processing and model development methods in the studies. Sub-figure a) Frequency of hyperparameter tuning and values reporting (reported in section 2.3.3). Sub-figure b) Proportion of hyperparameter tuning techniques (reported in section 2.3.3). Sub-figure c) Missingness reporting and imputation in papers (reported in section 2.3.2). Sub-figure d) Handling class imbalance in studies (reported in section 2.3.2).	15
2.5	Frequency of algorithms used on datasets with very small (fewer than 300), small (300-700), medium (701-2000), large (2001-20000) and very large (more than 20000) sample size.	16
3.1	The eRAPID trial design. It involved 1:1 randomization (stratified by cancer site, gender, previous chemotherapy) to allocate patients to usual care or eRAPID intervention groups.	26
3.2	The expected schedule of collecting PROMs and symptom severity reports in the eRAPID clinical trial.	28
3.3	A frequency plot of the number of time-steps (reports) completed.	33
3.4	Overview of Research Methodology.	34
3.5	Summary of the research methodology as it is developed in the thesis. Research questions were formulated to address the limitations of data and model type, design robustness, and stakeholder engagement. The chapters where the corresponding questions are addressed are shown. Time-series PROMs refer to both symptom severity reports and standardised questionnaires collected longitudinally.	39
4.1	Flow diagram illustrating the methodology of the study.	42
4.2	Number of non-missing data in each variable.	44
4.3	AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting admissions.	52

4.4	Calibration plots of LR models in original dataset (left), calibration of RF models in balanced datasets (middle), and calibration of LR models in partially balanced datasets (right). LR, logistic regression, RF, random forest.	53
4.5	AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting triage.	54
4.6	AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting changes to chemotherapy.	55
5.1	Flow diagram illustrating the overall methodology of the study. . . .	65
5.2	AUC values of models predicting A) physical well-being, B) social functioning, C) usual activities, D) mobility deterioration and improvement at 6, 12, and 18 weeks.	74
6.1	Flow diagram illustrating methodology in this chapter. Four models were applied on data collected throughout eRAPID clinical trial to predict outcomes during chemotherapy. Nested cross-validation was used to develop and evaluate models. Models built on data including and excluding PROMs were compared.	80
6.2	Balanced accuracy, AUC, specificity and recall plots for hospital utilisation outcomes predicted by all models.	87
6.3	Balanced accuracy, AUC, specificity and recall plots for HRQoL outcomes predicted by all models.	89
7.1	Flow diagram illustrating data flow and the nested cross-validation (CV) pipeline for implementing LSTM. The methodology for implementing ML models was described in Chapter 6.	98
7.2	Flow diagram illustrating the LSTM model. Time-series input is processed through LSTM layer, which output is then passed through fully connected layer, receiving separate static input. The model hyperparameters were established during nested cross-validation process. . . .	99
7.3	Balanced accuracy values for LSTM and ML predictions of all outcomes (plotted on the right side).	103
7.4	SHAP plots indicating feature importance for short- and long-term admission predictions. ROL - role functioning, SOC - social functioning, FWB_overall - functional well-being, d_since_cycle - days since last chemotherapy cycle, EWB_overall - emotional well-being, d_since_previous - days since previous report, SWB_overall - social well-being, ed_lev - level of education, EQ5DUsuAct - usual activities, PrimaryorMet - primary or metastatic, DCmarital - marital status, EQ5DPain - pain, EQ5DAnxDep - anxiety and depression, PreviousChemo - previous chemotherapy (yes,no), SOB - shortness of breath, EQ5DMob - mobility, EQ5DSelCar - self care	104

7.5	SHAP plots indicating feature importance for physical well-being and usual activities predictions. ROL - role functioning, SOC - social functioning, FWB_overall - functional well-being, d_since_cycle - days since last chemotherapy cycle, EWB_overall - emotional well-being, d_since_previous - days since previous report, SWB_overall - social well-being, ed_lev - level of education, EQ5DUsuAct - usual activities, PrimaryorMet - primary or metastatic, DCmarital - marital status, EQ5DPain - pain, EQ5DAnxDep - anxiety and depression, PreviousChemo - previous chemotherapy (yes,no), SOB - shortness of breath, EQ5DMob - mobility, EQ5DSelCar - self care	105
8.1	Aspects discussed with stakeholders throughout the study design, and the impact stakeholder engagement had on each part of methodology.	118
A1	PRISMA checklist for scoping reviews part 1 (Chapter 2).	233
A2	PRISMA checklist for scoping reviews part 2 (Chapter 2).	234
A3	TRIPOD+AI checklist part 1 (Chapter 3).	238
A4	TRIPOD+AI checklist part 2 (Chapter 3).	239
A5	All results obtained from the ML models in Chapter 4.	240

List of Tables

2.1	Inclusion and Exclusion criteria for the study selection	10
2.2	Machine learning and deep learning models used on data collected in one and multiple timepoints, ordered by number of publications. . . .	17
2.3	Methods of addressing temporality in time-series data, with sample size and frequency of self-reports. Three papers which used time-series data did not report how the temporality is addressed [142, 165, 166], and are not included in this table.	17
3.1	Examples of questions from the symptom severity reports from eR-APID clinical trial.	28
3.2	Data types in the dataset used in the research reported in this thesis.	30
3.3	Summary of stakeholders involved in the the design of research reported in this thesis.	35
4.1	Variables in each variable sets. The highlighted bold variables were selected to be included in the models, and the light variables were removed based on the high correlation coefficient.	47
4.2	Differences in variables between classes in all target variables and the p-value (* $P \leq .01$, ** $P \leq .001$, *** $P \leq .05$) generated by Mann-Whitney U test for all variables but AgeStudyEntry (t-test used). Mean (M), counts of values for categorical variables, and mean for continuous variables are reported.	48
4.3	Variable importance for LR (coefficient) and RF (value) models predicting all 3 target variables. The lower the rank of the variable, the higher the importance for a particular model.	57
5.1	Predicted variables, questionnaires completed, change calculations, participants number (N), including deterioration (D) and improvement (I) cases.	66
5.2	Removed variables based on correlation analysis.	67
5.3	Names of all variables used in the research presented in this chapter and the number of missing data in each variable for each outcome at 18 weeks.	68
5.4	The results of HRQoL deterioration prediction at each time-point. .	72
5.5	The results of HRQoL improvement prediction at each time-point. . .	73
6.1	Class distribution in each hospital utilisation and HRQoL outcome . .	83
6.2	Results of all models predicting hospital utilisation and HRQoL changes outcomes.	86

6.3	Comparison of results of models trained on data including and excluding PROMS predicting admission and admission within 14 days. .	88
6.4	Number of features within different significance levels of the difference between classes for each outcome.	92
7.1	Results of LSTM models developed in this chapter and ML models developed in chapter 6 predicting all outcomes of hospital utilisation and HRQoL changes.	102
8.1	The framework for developing and evaluating clinically relevant AI models applied on PROMs.	120
A1	Search strategy for the scoping review (Chapter 2).	235
A2	Study characteristics and pre-processing methods used by studies included in the review (Chapter 2).	236
A3	Model development and evaluation, including study characteristics of papers included in the review (Chapter 2).	237
A4	Hyperparameters of each model predicting deterioration of the HRQoL outcomes (Chapter 5).	241
A5	Hyperparameters of each model predicting improvement of the HRQoL outcomes (Chapter 5).	242
A6	Feature importance ranks for physical well-being changes at 18 weeks prediction models LR and RF (Chapter 5).	243
A7	Feature importance ranks for usual activities changes at 18 weeks prediction models LR and RF (Chapter 5).	244
A8	Feature importance ranks for mobility changes at 18 weeks prediction models LR and RF (Chapter 5).	245
A9	The number of missing data (out of 3260) per column (Chapter 6). .	246
A10	Variable importance os LR and RF predicting admission and admission within 14 days from a symptom report (Chapter 6).	247
A11	Variable importance os LR and RF predicting deteriorations in physical well-being and usual activities during chemotherapy (Chapter 6). .	248

List of Acronyms

AB Adaptive Boosting

AI Artificial Intelligence

AUC Area Under the Receiver Operating Characteristic Curve

ANOVA Analysis of Variance

BA Balanced Accuracy

BMI Body Mass Index

CC Complete Case

CNN Convolutional Neural Networks

CONSORT-PRO Consolidated Standards of Reporting Trials Patient Reported Outcomes Extension

CTCAE Common Terminology Criteria for Adverse Events

CV Cross-Validation

DL Deep Learning

DT Decision Tree

EHR Electronic Healthcare Records

EORTC European Organisation for Research and Treatment of Cancer

EORTC QLQ-C30 European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire

EQ-5D EuroQol 5 Dimensions 5 Level

eRAPID Electronic patient self-Reporting of Adverse-events: Patient Information and aDvice

FACIT Functional Assessment of Chronic Illness Therapy

FACT-G Functional Assessment of Cancer Therapy-General

FDA U.S. Food and Drug Administration

HRQoL Health-Related Quality of Life

IJMEDI International Journal of Medical Informatics

KNN K-Nearest Neighbours

LR Logistic Regression

LSTM Long-Short Term Memory

MCAR Missing Completely at Random

MCID Minimally Clinically Important Difference

ML Machine Learning

MLP Multilayer Perceptron

NHS National Health Service

PCOR Patient Centred Outcomes Research

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses

PRO-CTCAE Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events

PROMs Patient-Reported Outcome Measures

PROs Patient-Reported Outcomes

PROTEUS Patient-Reported Outcomes Tools: Engaging Users and Stakeholders

RF Random Forest

RNN Recurrent Neural Networks

SHAP Shapley Additive Explanations

SISAQOL Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints

SMOTE Synthetic Minority Over-sampling Technique

SVM Support Vector Machines

TRIPOD+AI Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis adapted for machine learning research

Tukey HSD Tukey's Honest Significant Difference

WHO World Health Organisation

XGBoost Extreme Gradient Boosting

Glossary

Adaptive Boosting (AB) - An ensemble machine learning classification model combining different classifiers to complement their weak points[6].

Artificial Intelligence (AI) - A field of computer science and engineering which uses computer systems able to mimic intelligent behaviour [7]. AI is a very broad term, which can involve interpreting, generating and processing text, speech, images, or numerical data [8]. This thesis focuses on AI prediction models which can learn from available data to predict the likelihood of outcomes [9]. Types of AI models used in this thesis are differentiated between Machine Learning (ML) and Deep Learning (DL) to explore the potential of different models for processing patient-reported data:

1. Machine Learning (ML) - simple AI models with a single layer [10].
2. Deep Learning (DL) - it is a subset of machine learning, which refers to more complex neural network architectures with a large number of layers and parameters [11].

Class imbalance - A situation where two classes of an outcome in binary classification task have unequal number of cases. It is a common problem in AI research, as most models do not learn well from the imbalanced data [12].

Cross-Validation (CV) - The process of dividing the dataset randomly into x number of folds, and refitting the model x times. It can be used for hyperparameter tuning or aggregating the performance of the models from each fold to avoid bias resulting from a random train/test split. Exploration of hyperparameter tuning technics can help understanding how the models were developed.

Data leakage - In AI modelling, data leakage is an accidental sharing of information between train and test sets which can result in performance estimates being better than in the intended application scenario [13].

Decision Tree (DT) - A model which uses nodes and branches to perform splitting through choosing between input variables based on a selected parametric criterion [6].

Extreme Gradient Boosting (XGBoost) - A machine learning algorithm that works based on gradient boosted decision trees [14].

Health-Related Quality of Life (HRQoL) - A subjective perception of

the impact of health (including disease and treatment) on physical, mental, social functioning [15]

Hyperparameter tuning (also known as hyperparameter optimisation) - a technique for the selection of the best hyperparameters to enhance performance of AI models [16]. The model with hyperparameters which provide the best performance in training is then used for evaluation [17].

Logistic Regression (LR) A binary classifier estimating regression equation for a target variable [6].

Long-Short Term Memory (LSTM) - recurrent neural network. It is applied to time series or sequence data for outcome prediction [18]

Multilayer Perceptron (MLP) - A model that mimics human brain through interconnected neurons. It includes an input layer, hidden layers, and an output layer [19].

Patient-Reported Outcomes (PROs) - Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. [20]

Patient-Reported Outcome Measures (PROMs) - Instruments used to collect PROs, including disease related symptoms, side effects of treatments, HRQoL, and impact on functioning. PROMs data can be collected either through standardised and validated tools designed to capture patients' reports, or any other forms of symptom and quality of life measures [21].

Random Forest (RF) - Another ensemble classifier using decision trees on various generated train sets and uses averaging to improve the predictive performance [6].

Support Vector Machines (SVM) - A model that uses complex variable space transformation to discriminate between classes applying a hyperplane [19].

Chapter 1

Introduction

1.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) is a field of computer science and engineering which uses computer systems able to mimic intelligent behaviour [7]. AI is a very broad term, which can involve interpreting, generating and processing text, speech, images, or numerical data [8]. With the recent advances in technology, AI has become a part of everyone's lives. For example, virtual assistants play a huge role in customer services, running of businesses, and even providing healthcare advice. Furthermore, AI-enabled robots can automate variety of tasks, from household chores to medical surgeries [22].

This thesis focuses on the type of AI which can learn from available data to predict the likelihood of outcomes [9]. In data science, drawing insights from data and driving decision-making progress has been originally addressed by classical statistics. This involves obtaining a mathematical model to process available data and determine what is likely to happen in the future [23]. AI still involves statistics and calculations, but it attempts to simulate human reasoning in performing prediction tasks [24]. The difference between statistical modelling and AI predictive models is unclear, causing confusion in scientific community [25]. Friedrich et al., suggested that AI relies on a dataset during learning and evaluation, whilst statistics use theory, which provides a prior for making assumptions [26]. AI predictive models are known to be more helpful than classical statistical methods when applied to larger and more complex datasets, as they can identify patterns in high-dimensional environments [27].

1.2 Artificial Intelligence in Medicine

AI has revolutionised many industries, such as finance, entertainment, or transportation. It is also increasingly common to utilise AI in healthcare [28]. The field of medicine has been struggling with serious issues, such as diagnostic and treatment errors, inequalities in care, operational inefficiencies, and limited time for patient-clinician engagement [29]. The need for improvements in medical practice and the growing volume and complexity of healthcare data provide the opportunities for AI to be implemented [30]. Accurate and robust AI systems can process patients' information from large datasets to help personalise treatments,

improve timely diagnosis, minimise human errors, and as a result save time and resources [31]. These can have a positive impact on patient outcomes, which include symptoms of disease, adverse effects of medical interventions, survival, and quality of life [32].

AI models predicting patient outcomes have been applied to various data types in medicine, such as Electronic Healthcare Records (EHR), diagnostic images, genetic data, or brain activity data [33]. For example, EHR data were used to predict readmission and mortality risks, which has a potential to plan resource allocation and prevention of adverse events [29]. Furthermore, AI systems can process radiology images [34] and digitalised biopsies [35] for early cancer detection. AI has also been utilised in mental health, predicting successful medication, suicide risk and diagnosis of depression [29].

Despite the growing potential for applications of AI in clinical practice, there are many challenges which limit implementation opportunities. These include poor quality of medical datasets, the focus on providing new solutions rather than validating existing models, and regulatory issues [34]. Explainability of AI systems is also very important to understand why a model is making certain decisions, but many AI models are extremely difficult to interpret [29]. Furthermore, medical datasets used for training and evaluating AI models are rarely representative of the population. This suggests that using AI in medicine can further deepen healthcare inequalities and under-perform for minority groups [36]. Another important challenge is a lack of public trust in AI. Patients and clinicians are rarely involved in the design of studies using AI in healthcare, which limits the feasibility of explored solution and stakeholders' understanding of how AI systems can help address a healthcare need [37].

1.3 Patient-Centred Care

Patients' and clinicians' views about a clinical encounter sometimes differ. For instance, patients may believe that they were involved in the decision-making process, yet often the observation of clinician-patient interactions during appointments suggests low level of patients' participation [38]. Listening to patients' views on their symptoms and understanding their expectations have the potential to reduce anxiety and as a result decrease healthcare utilisation and costs [39]. Patient-centred care is a concept which assumes treating patients as equal partners [40]. It can enable patients to manage and improve their health and personalise their experiences in receiving care. Patient-centred care also prevents dependency and underestimating patients' abilities to protect their own health [40].

Unfortunately, there are many barriers to the implementation of patient-centred care in practice, as described by Kwame et al., [41]. These include the lack of time and resources to sufficiently interact with patients. The healthcare units also provide environmental barriers, such as lack of privacy, noisy surroundings, and poorly coordinated services. Shortage of staff, high demand and burnout often prevent effective communication and personalised approach to care. Furthermore, personal and behavioural barriers, such as cultural background and beliefs can result

in discrimination. Patient-centred care can be achieved by changing the dynamic of clinical appointments and focusing on quality of life. It can be improved through understanding patients' views on their physical, social and mental well-being [41].

1.4 Patient-Reported Outcome Measures (PROMs)

1.4.1 History of PROMs

In 1952, World Health Organisation (WHO) defined health as the presence of positive physical, mental and social well-being, rather than the absence of disease or infirmity [42]. The definition was timely, as the focus of medical care moved from treating acute diseases (such as infectious diseases) to managing long-term chronic conditions (such as diabetes, asthma and chronic lung disease). This change in patterns of diseases shifted the health-related thinking and increased the interest in measuring overall well-being. The first questionnaires of health and functional status were created to assess functioning and well-being in large populations (Rand Medical Outcomes Study [43], Sickness Impact Profile [44]). These questionnaires were useful to describe populations well-being but they did not focus specifically on concepts of interest.

The idea of measuring physical, social and mental well-being was further developed into Health-Related Quality of Life (HRQoL) conceptual model [15]. It is defined as “subjective perception of the impact of health (including disease and treatment) on physical, mental and social functioning”. Instruments measuring HRQoL were multidimensional and specifically developed to measure outcomes in clinical trials. For example, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) [45] and Functional Assessment of Cancer Therapy-General (FACT-G) [46] have been widely used in cancer trials, usually as secondary outcome measures.

In early 2000, the concept of measuring patients' perception on their own health status further emerged to Patient-Reported Outcomes (PROs). They are defined by the U.S. Food and Drug Administration (FDA) as “any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else.” [20]. Patient-Reported Outcome Measures (PROMs) are instruments used to collect PROs, including disease related symptoms, side effects of treatments, HRQoL, and impact on functioning [47]. Validated and standardised PROMs questionnaires (including HRQoL) are often used in clinical trials for the collection of outcome measures. The data collection has been usually completed on paper (in clinics or postal questionnaires) or through interviews.

The methodology for implementing PROs in trials has been developing and maturing, resulting in a range of international consensus guidelines, such as Consolidated Standards of Reporting Trials Patient Reported Outcomes Extension (CONSORT-PRO) [48], Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension [49], the Patient-Reported Outcomes Tools: Engaging Users and Stakeholders (PROTEUS) Consortium [50],

and Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints (SISAQOL) [51]. Furthermore, recommendations for ethical issues that should be addressed in clinical research involving PROs are included in the PRO Ethics Guidelines [52]. Another important guidance was provided by FDA through a core set of PROs [53], which included disease-related symptoms, symptomatic adverse events, overall side effect impact summary measure, physical function and role function. If there is any other aspect that needs to be measured, FDA recommends careful consideration of patient burden and the quality of data collected by focusing on the most meaningful and measurable outcomes.

One limitation of the standard validated questionnaires is the fixed structure and the lack of flexibility to report relevant symptoms. For example, cancer treatments have constantly changed with implementation of new drugs (such as immunotherapy, antibodies and other targeted therapies), with a wider range of toxicity than traditional chemotherapy [54]. To address the need for more flexible measurements of symptoms and HRQoL in cancer trials, the use of item lists from item libraries that contain all available items (questions) derived from established measurement systems (such as European Organisation for Research and Treatment of Cancer (EORTC) [55] or Functional Assessment of Chronic Illness Therapy (FACIT) [56]), has been advocated [57]. Item libraries use interactive online platforms that allow the creation of customised questionnaires (item lists). This enabled flexible use of PROMs reports, depending on the measurement need.

One specific example of an individual item library is a Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) [58]. Common Terminology Criteria for Adverse Events (CTCAE) is a system which provides a terminology for reporting and grading of the severity of adverse events by oncologists in cancer clinical trials [59]. The CTCAE is completed by a healthcare professional, based on a consultation with patients. This system has been adapted by several research groups into a patient self-reporting version [60, 61], and most recently to National Cancer Institute's PRO-CTCAE.

1.4.2 Benefits of PROMs

There is an increasing drive to utilise solutions to collect patient-reported data during clinical trials. PROMs have been used as outcome measures in research, which provide an opportunity to understand and measure aspects of symptoms, functioning and HRQoL [55]. They can help establish the efficacy and tolerability of therapies through reported information. PROMs can also improve the interpretation of clinical informations, through understanding the impacts of illness of treatment, which patients are experiencing. Furthermore, the collection of PROMs, especially in longitudinal settings, provide high-value datasets which can further be used for analysis. Electronic collection has further facilitated PROMs use in clinical trials. It can reduce response burden, make administrative tasks easier and increase adherence to reporting [62].

In clinical practice, mobile applications have been widely used in healthcare and have the potential to improve the quality and personalisation of patient care

[63]. Electronic collection of PROMs can be more effective, time-efficient and minimize costs, compared to paper forms [64]. The information collected through electronic PROMs could support shared exploration of treatment options through in-depth analyses of current health state and quality of life expectations [65]. There are several studies exploring the benefits of PROMs in patient care. Evidence suggests that the use of electronic PROMs during cancer treatment can result in better patient well-being [66, 67]. Furthermore, the use of weekly electronic PROs to monitor symptoms resulted in improvements in physical functioning, symptom control, and HRQoL at 3 months during cancer treatment [68]. Importantly, the use of PROs in the routine care of patients with metastatic cancer was associated with increased survival compared with usual care [68, 69]. The systematic use of PROMs can improve communication and decision-making between doctors and patients, which leads to better patient satisfaction with care, and increased patient confidence in self-management. Furthermore, patients' level of independence, understanding of the health issue and responsibility for the management of the condition tend to increase after they start measuring and reporting symptoms [70, 71]. The inclusion of PROMs in clinical practice can support the understanding of the impact of treatment or a service on a patient and assess problems relevant and meaningful to patients.

1.5 PROMs in AI Research

The collection and accessibility of PROMs in clinical practice or clinical trials generate more data on patients' health status. This adds another element to datasets used for data analysis and provides more variables, which could be utilised in prediction models. Currently, AI models predicting patient outcomes rarely learn from any information gathered from patients about their own health and well-being. This suggests that the models process clinical and demographic information, without consideration of patients' perspective [72]. However, PROMs are increasingly used in research aiming to predict outcomes in medicine with statistical methods, or more advanced technologies, such as AI [73]. PROMs provide the representation of patients' views, which can be utilised in AI medical research. A systematic evaluation of PROMs in clinical trials of AI health technologies has shown that patients' perspective is central even in novel technological advancements [74].

PROMs are utilised either as input or output data in AI models. The input data, which consists of various variables used as predictors, usually include PROMs alongside other types of data, for example clinical and demographic. Including PROMs as input variables enables AI models to process subjective information from patients, which no clinical or demographic measures can provide. This can help increasing clinical relevance of models, and as a result, often improve model performance [6]. However, PROMs can also be used as output data. Usually, the predicted variables involve the Minimally Clinically Important Difference (MCID) of a particular questionnaire. For example, studies in orthopaedics often calculate the difference between pre- and post-operative data collection. This difference becomes a target variable to establish efficacy of the intervention [75]. This thesis explores the use of PROMs as both input and target variables.

Despite the increasing number of research using PROMs as predictors in AI models foreseeing health outcomes [73, 76], there is a lack of standard approach for handling the complexities of PROMs data, which include irregular reporting, missing data, and typically small datasets [72]. These challenges lead to inconsistent approaches for data pre-processing, which could bias results and affect the interpretation of model performance. The lack of universal guidelines can limit the understanding of overall potential of AI applied to PROMs data, as the comparison of model performances across different papers is less meaningful when different methodologies are used [77]. Furthermore, research often lacks reproducibility, external validity [78], and generalisability of the results to the clinical context [79]. Inadequate and inconsistent selection of PROMs also introduces a challenge to useful evaluation of patient-centric AI models in healthcare [72]. The research reported in this thesis will focus on understanding the predictive value of different types of PROMs data in AI models, identifying approaches for data pre-processing and model development which can mitigate bias, as well as involving stakeholders in research design.

1.6 Aims and objectives

There is limited research exploring and addressing methodological gaps of applying AI on PROMs data to predict patient outcomes. The lack of guidelines on utilising PROMs in AI predictive models and inconsistencies in research design limit the potential of such tools to be implemented in clinical practice [73]. This thesis proposes a novel comprehensive evidence-based framework for designing research applying AI on PROMs data to predict patient outcomes. It will be achieved through literature review and original research. The literature review will help understand research gaps in the literature, which should be addressed to improve the quality of research using PROMs and AI. Furthermore, this thesis will use a dataset which consists of longitudinal PROMs, clinical and demographic information, collected throughout a clinical trial from patients undergoing chemotherapy, to propose methodological approaches for using PROMs and AI to predict patient outcomes. These approaches will be derived through research meeting the following objectives:

- To explore the predictive potential of PROMs in AI predictive research. This objective will help understand if adding information collected from patients to clinical and demographic variables improves model performance.
- To apply and compare the most common methodologies for pre-processing various types of PROMs data and assess their effects on prediction performance of AI models.
- To apply patient-centric design of research, which includes stakeholder engagement to understand clinical relevance of AI research.
- To develop state-of-the-art AI models maximising the potential of PROMs collected at single time-point or multiple time-points in predicting patient outcomes.

1.7 Thesis structure

This thesis consists of eight chapters in total. This introductory chapter has provided a brief background on the potential of using PROMs in AI predictive models, as well as the aims and objectives of this thesis. Chapter 2 presents a scoping review which explores the approaches used in literature for utilising PROMs and AI for patient outcome predictions. It establishes research gaps in the literature which this thesis addresses through original research. Chapter 3 outlines the methodological choices of the research reported in this thesis. It describes the dataset used and provides the overview of all phases of the research. Chapters 4 - 7 are primary research chapters which aim to meet the objectives outlined in the previous section. They utilise a dataset which comes from a clinical trial in the cancer setting, where PROMs were collected as part of routine care symptom monitoring and wider outcome assessment. These chapters explore the predictions of patient outcomes using AI and PROMs through the following examples:

- Chapter 4 uses PROMs collected at one time-point prior to starting chemotherapy to predict hospital utilisation during chemotherapy.
- Chapter 5 uses PROMs collected at one time-point prior to starting chemotherapy to predict HRQoL changes during chemotherapy.
- Chapters 6 and 7 focus on assessing different types of AI models using PROMs collected longitudinally throughout chemotherapy to predict hospital utilisation and HRQoL changes during chemotherapy.

Chapter 8 is a conclusion chapter, which summarises the outcomes of this research and describes the proposed framework for utilising PROMs and AI for predicting patient outcomes. It clearly outlines the contributions which this thesis brings to the field of AI in medical research.

Chapter 2

Literature Review

2.1 Introduction

This chapter is an important foundation for the thesis to help understand the current status and breadth of existing research using PROMs in AI predictive modelling and the methodological approaches used. It presents a scoping review, which helps identify research gaps in literature and develop research questions, which are addressed in this thesis. This scoping review aims to investigate methodological approaches of primary studies in a variety of healthcare domains which used AI for predicting patient outcomes using PROMs. It can support the understanding of the main challenges of including PROMs in AI predictive models. Furthermore, it can identify opportunities for improvement of the quality of such research.

There are existing literature reviews exploring AI models applied on PROMs data. For example, a scoping review from 2021 investigated PROMs as stand-alone input variables in models. However, the added value of PROMs in predictive models is not analysed, as the analysed studies did not include any clinical or demographic data. Furthermore, the reproducibility and clinical adoption of studies are not explored in depth. The literature search included only 2 medically-oriented literature databases, without involving databases from engineering and computer science backgrounds [73]. Other existing reviews have focused on selecting individual healthcare domains, but have not investigated all potential applications of methods applied on self-reported data in a wide range of fields of medicine [80, 81].

This review considers primary research from various healthcare domains, which used PROMs data as a stand-alone input, or together with other types of data for prediction of patient outcomes. Six different databases are searched, enabling exploration of medical, engineering, and interdisciplinary literature. The review focuses on the following aspects of applying AI methods to PROMs:

- The opportunities and challenges of including PROMs in AI models predicting patient outcomes.
- Methodological rigour and reporting of pre-processing PROMs data prior to applying AI models.
- Pipelines used to develop and evaluate AI models using various types of PROMs

- Clinical relevance of developed models, including stakeholder engagement and readiness to be applied in clinical practice.

The final database search of this review is conducted on the 7th November 2023 and the review is published in Health and Quality of Life Outcomes journal on the 11th April 2025 [1]. The conclusions from this review guided the methodology of this thesis. Therefore, the most recent studies have not been included in the analysis. A brief overview of recent literature which is not included in the review is written in the Discussion section (2.4.6) of this chapter.

2.2 Methods

The methodology of this scoping review is based on the Joanna Briggs Institute guidance [82]. The review protocol is available on Open Science Framework [83]. The completed Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) checklist for scoping reviews [84] is added in the Appendix Figures A1 and A2.

2.2.1 Search Strategy

The databases used to search for relevant papers are: Web of Science, IEEE Xplore, ACM Digital Library, Cochrane Central Register of Controlled Trials, Medline and Embase. The keywords adapted to each database are listed in Appendix Table A1. Initially, the limited search of Web of Science and Medline is conducted to analyse and approve the keywords. The finalised search of all the records across all the databases is completed on the 7th November 2023. The search strategy with all keywords adapted for each database is provided in the Appendix Table A1. The reference lists of all relevant reports are also screened. All studies identified through the search strategy are exported to Endnote citation management system.

2.2.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria follow the Population/Concept/Context (PCC) framework [82] and are described in Table 2.1. The participants in the papers included in the review are patients, whose symptoms and quality of life data are recorded using various PROMs. These can include mobile applications, or surveys completed either online or in a clinic. The type of data can be collected through either validated and widely used PROMs, or any other patient self-reports. Papers reporting the use of PROMs data as both predictive and predicted variables are included in the review. If PROMs data are solely used as a predicted variable, and not included as inputs, the reference is excluded, as the review will focus on data pre-processing and model development. The concept is the methods of AI used for predictions of patient outcomes. Papers that explicitly mentioned use of AI or Machine Learning methods are included. Any papers using AI models for purposes other than prediction are excluded. Papers that reported prediction of patient outcomes in the healthcare context are included in the analysis. These outcomes should belong to the categories of patient-related outcomes identified by Kersting et al (2020) [32], presented in Table 2.1. These categories provide the

broad understanding of healthcare context, which allows focusing on the AI used for patient-relevant reasons.

2.2.3 Review Process

All duplicates found in the databases are removed automatically in Endnote. Titles and abstracts of the papers are screened by a researcher and re-selected based on inclusion and exclusion criteria, presented in Table 2.1. Full texts of articles admitted to the review are assessed against the inclusion and exclusion criteria again. The researcher's approach is validated through a second reviewer, who repeated scanning through 10% of abstracts and selected full texts, and compared their results with the first reviewer. The validation showed high consistency between the reviewers' decisions, as out of 218 validated papers, 184 (84.4%) are consistently selected or rejected. Therefore, no further validation is performed.

Table 2.1: Inclusion and Exclusion criteria for the study selection

Category	Inclusion criteria	Exclusion criteria
Input variables	Data reported by patients using standardised PROMs; electronic data collection designed for self-reporting of symptoms or clinical outcomes. These can be used with combination of different non-patient reported data.	Only not patient-reported data, e.g., data reported through a clinician, recorded, written down during an appointment, data reported on online forums/social media, or data from physiological measurements.
Output variables	Any data describing patient outcomes: sleep behaviour, coping and self-efficacy, healthcare utilisation, body image perception, function, communication skills, reliability of diagnosis and therapy, optimal support, confidence in therapy, satisfaction, cognitive performance, treatment decision, disease control, daily activities, reoperation, mental health, quality of life, mobility, co-morbidities, pain, survival, adverse events, and symptoms.	Output that does not relate to health or healthcare, or papers which did not aim to predict outcomes.
Models	Machine learning or deep learning predictive models.	Statistical models (e.g., only regression analysis).
Paper type	Primary research reported in English language	Abstracts only, theses, dissertations, letters to editors, guidelines, commentaries, introductions, papers published not in English, or review papers.

2.2.4 Data Extraction and Analysis

The data are extracted from all papers selected for this review. Extracted information for each included paper is presented in Appendix Tables A2 and A3. A second reviewer extracted data from 10% of admitted papers for the purpose of validation, and the extracted information is compared and agreed between the reviewers. The summary of data analysed is reported in the narrative and tabular form. The analysed data include:

1. Study characteristics:

- Country and year of paper publication - to establish trends of using AI for predicting patient outcomes.
- Healthcare domain (e.g., oncology, orthopedics) - to explore the fields of medicine which utilise PROMs and AI research.
- Types of PROMs questionnaires used, and whether the collection is completed only at one time-point (static data) or across different time-points (time-series data) - to explore how different types and temporality of PROMs data can be processed in AI research.

- Sample sizes - to understand the potential of using datasets of different sizes in the AI research.

2. Data pre-processing methods:

- Missingness in the datasets and missing data imputation techniques - to explore how incomplete datasets have been handled.
- Class imbalance - a situation where two classes of an outcome in binary classification task have unequal number of cases. It is a common problem in AI research, as most models do not learn well from the imbalanced data [12]. Techniques for handling class imbalance will be explored to understand how effects of the imbalance can be mitigated.

3. Model development:

- Types of AI models used - differentiated between ML and DL to explore the potential of different models for processing PROMs. Even though DL is a subset of ML, this chapter, as well as the rest of the thesis will differentiate between them for the purpose of defining simpler and more complex models:
 - (a) Machine Learning (ML) - in this thesis, ML will be referred to as simple AI models with a single layer [10].
 - (b) Deep Learning (DL) - More complex neural network architectures with a large number of layers and parameters [11].
- Hyperparameter tuning (also known as hyperparameter optimisation) - a technique for the selection of the best hyperparameters to enhance performance of AI models [16]. The model with hyperparameters which provide the best performance in training is then used for evaluation [17].
- Cross-Validation (CV) - The process of dividing the dataset randomly into x number of folds, and refitting the model x times. It can be used for hyperparameter tuning or aggregating the performance of the models from each fold to avoid bias resulting from a random train/test split. Exploration of hyperparameter tuning technics can help understand how the models are developed.

4. Model evaluation:

- Performance metrics used - to understand how the performance of models is evaluated.
- Variable importance - to identify which factors have a potential to have an effect on predictions.
- Predictive value of PROMs - to explore whether or not including PROMs improved the model performance.

5. Clinical relevance and adoption:

- Patients and clinicians involvement in the study design - to establish if the approach is patient-centric and addressed a real clinical need.

- Stage of research - to explore if the validation and deployment of the AI models have been done.
- Reporting of sociodemographic information - to understand the potential bias which might arise from used datasets.

2.3 Results

Out of 3077 records screened, 94 are selected for analysis in this review. PRISMA diagram [84] (Figure 2.1) illustrates the process of paper selection. The reasons for paper exclusions are: no full-text available (38.1%), no PROMs used as input variables (35%), or no AI models used (12.%), or methods did not aim to predict patient outcomes (8.75%).

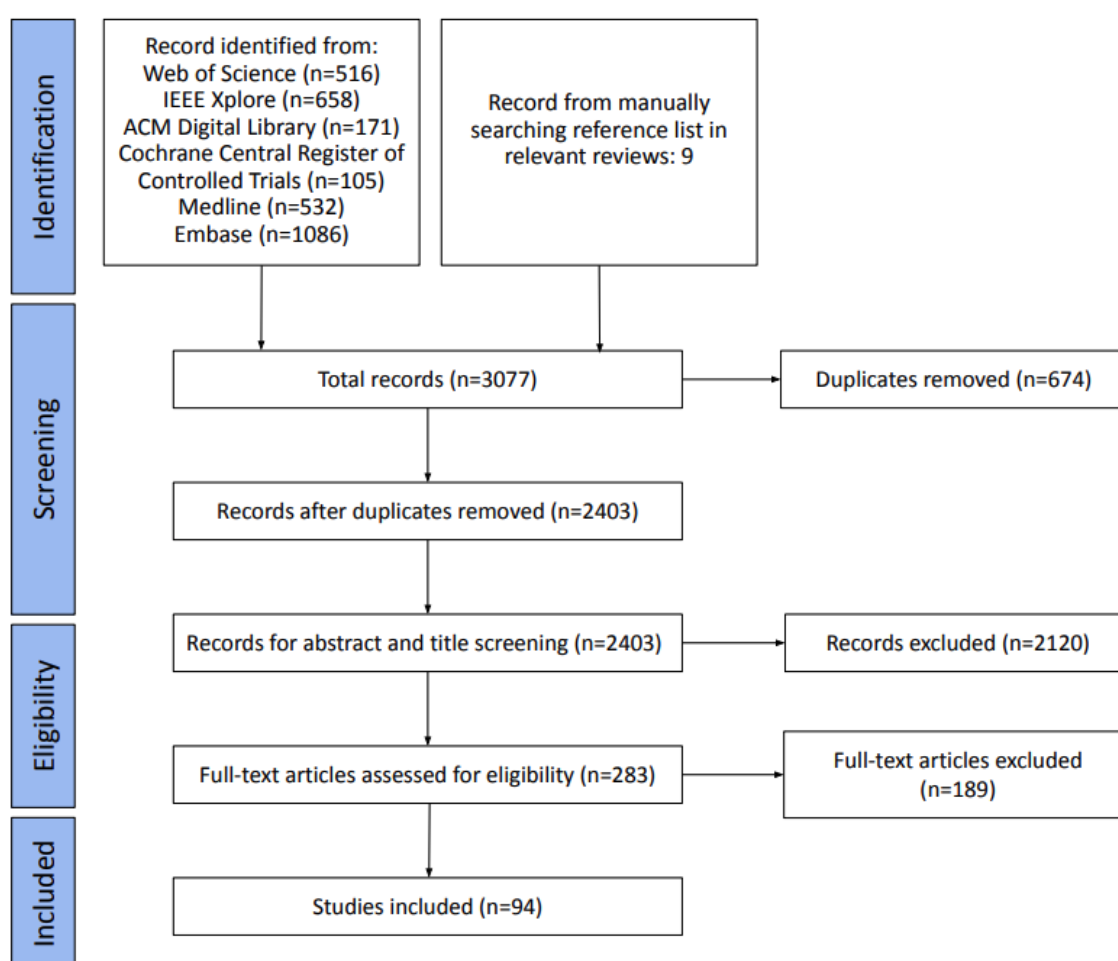


Figure 2.1: PRISMA flow diagram.

2.3.1 Study Characteristics

Among the identified studies, 33 (35%) are conducted in USA, 31 (33%) in Europe, 6 (6%) in Canada, 15 (16%) in Asia, 1 (1%) in South America, 1 (1%) in New Zealand, and 1 (1%) in Turkey. Six studies are conducted internationally (USA and Canada (n=2, 2%); UK and USA (n=2, 2%); Canada and Sweden (n=1, 1%);

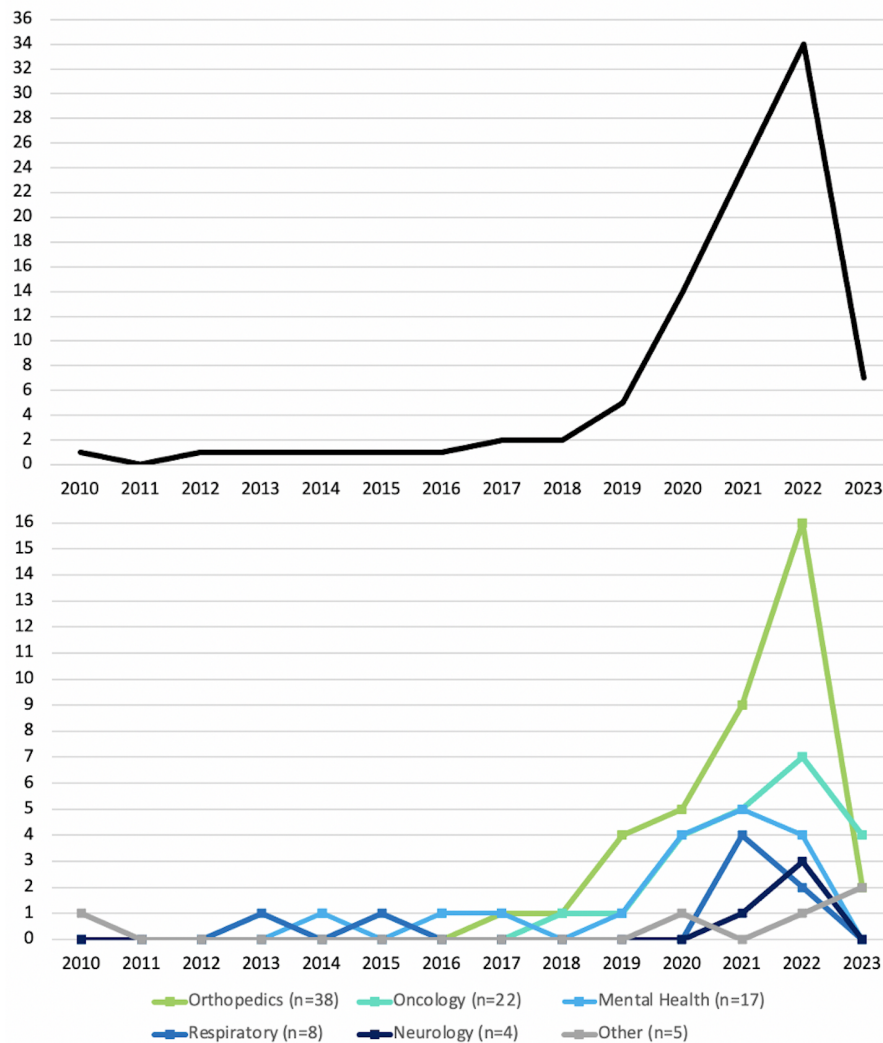


Figure 2.2: Year of publication of all 94 studies (top figure) and distribution of the studies based on their healthcare domain (bottom figure).

Europe, US, Australia and Israel ($n=1$, 1%). Identified papers focused on orthopedics ($n=38$, 40%), oncology ($n=22$, 23%), mental health ($n=17$, 18%), respiratory ($n=8$, 9%), neurology ($n=4$, 4%) and other domains, which appeared only once: hearing, endometriosis, palliative care, sub-health state, and cardiovascular. The studies found by the search strategy, which met eligibility criteria are published between 2010 and 2023 (Figure 2.2). The data are obtained either from existing registry/database ($n=44$), or pre-existing or current research studies ($n=47$, not reported: $n=3$). The self-reported input variables are combined with clinical and demographic data ($n=63$, 67%), only demographic data ($n=14$, 15%), only clinical data ($n=3$, 3%), or other types of data ($n=4$, 4%), such as wearable, electroencephalographic, bio-mechanical, or family data. Ten studies used only self-reported data for predictions. Most papers ($n=63$, 67%) are predicting self-reported outcomes, 14 (15%) of which are Minimally Clinically Important Differences (MCID) between pre- and post- clinical event data collection. Other papers used either only objectively measured (not self-reported) outcome ($n=28$, 30%), or a combination of self-reported and objective outcomes ($n=3$, 3%). Sample sizes of the papers varied from 20 to 1,434,868 (mean=25888, median=1022, 1st quartile=429.75, 3rd quartile=2879.75). The quartiles do not indicate clear boundaries between the data, as there are small differences in the sample sizes. Hence, a boundary-based approach is followed instead of quartile-based: very small (<300), small (300-700), medium (701-2000), large (2001-20000), and very large (>20000), as presented in Figure 2.3. The vast majority of studies ($n=69$) used condition-specific PROMs, such as orthopedic-specific Knee Injury and Osteoarthritis Outcome Score (KOOS) [85] or cancer-specific EORTC Core Quality of Life Questionnaire (QLQ-C30) [45]. In 31 papers condition-specific measures with generic questionnaires, for example EuroQol-5D (EQ-5D) [86]. Twelve papers used generic measures only. The questionnaires used focused on physical health ($n=18$), mental health ($n=11$), or both ($n=52$).

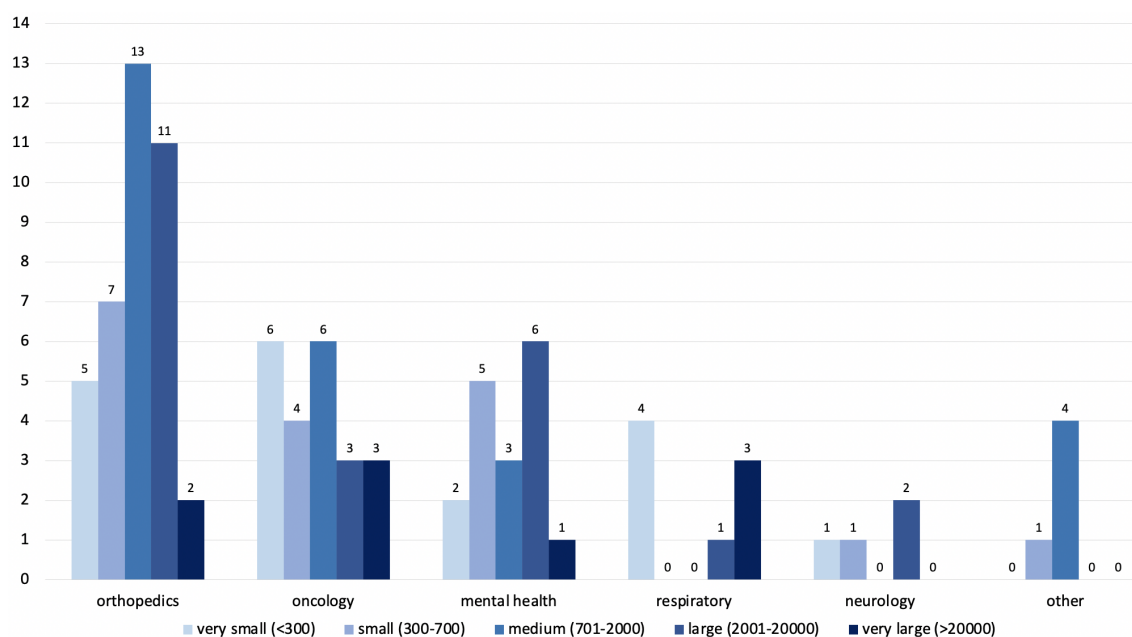


Figure 2.3: Number of papers categorised based on sample sizes and healthcare domain.

2.3.2 Data Pre-Processing

Thirty papers did not report the missingness in the dataset. It might mean that no missingness happened, or that missing data are not disclosed. Out of 64 papers which reported missing data, only 1 stated that there is no missingness in the dataset. Ten papers which reported having missing data did not report how the missingness is handled or addressed. Fifty-three papers (56%) reported the technique for data imputation (figure 2.4c)). The most common technique is complete case analysis ($n=16$) and mean/median/mode imputation ($n=15$). Most papers ($N=89$) used classification as a prediction method. Fourteen of these did not provide any information about the class distribution. Out of papers which did report class distribution ($n=75$, 80%), only 11 (15%) had balanced classes (maximum imbalance ratio of 60:40 between the minority and majority class [87]). Sixty-four papers (68%) used dataset with imbalanced classes, 29 (45%) of which did not mention the class imbalance problem. Thirty-five papers (55%) acknowledged the issue but 13 (37%) of them left the data imbalanced. Twenty-two papers reported the need for balancing the classes, but there is inconsistency in the methods across papers (Figure 2.4).

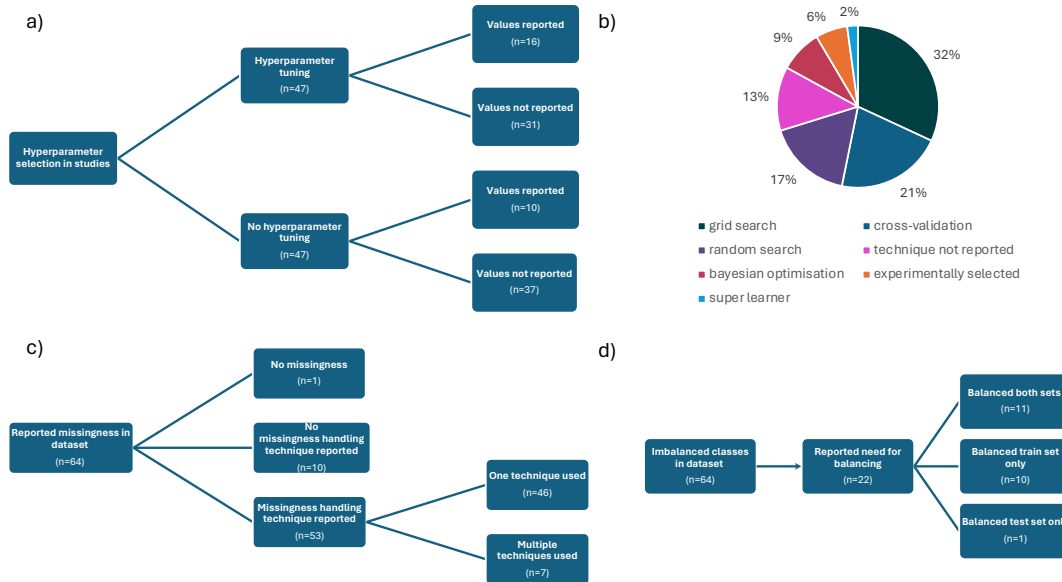


Figure 2.4: Reporting of pre-processing and model development methods in the studies. Sub-figure a) Frequency of hyperparameter tuning and values reporting (reported in section 2.3.3). Sub-figure b) Proportion of hyperparameter tuning techniques (reported in section 2.3.3). Sub-figure c) Missingness reporting and imputation in papers (reported in section 2.3.2). Sub-figure d) Handling class imbalance in studies (reported in section 2.3.2).

2.3.3 Model Development

Most papers ($n=84$, 89%) used multiple AI models for outcomes prediction. Forty papers (43%) used only traditional machine learning (ML) models, 5 (5%) only deep learning (DL) models, and 49 (52%) both ML and DL. The most frequently

used models are regression models ($n=61$, 65%), including linear, logistic, ridge and LASSO regressions; boosting methods ($n=53$, 56%), which included adaptive boosting, extreme gradient boosting and gradient boosting machine, random forest ($n=50$, 53%), artificial neural network ($n=43$, 46%), including single-or multi-layer perceptrons, and support vector machine ($n=39$, 41%) Figure (2.5). Most studies ($n=74$, 81%) applied AI models on data recorded in one time-point. The remaining studies trained their models on data collected in multiple time-points (Table 2.2). Out of these, 3 studies (3%) reported using models that process the temporal dependencies in the data, such as long-short term memory (LSTM) model [88, 89], and recurrent neural network with gated recurrent units (GRU) [90]. Nine studies considered temporality through coding it in the feature sets, and 5 papers did not address temporality at all (Table 2.3). Half of the papers ($n=47$, 50%) reported performing hyperparameter tuning, and out of these only 16 (34%) reported used hyperparameters (Figure 2.4a)).

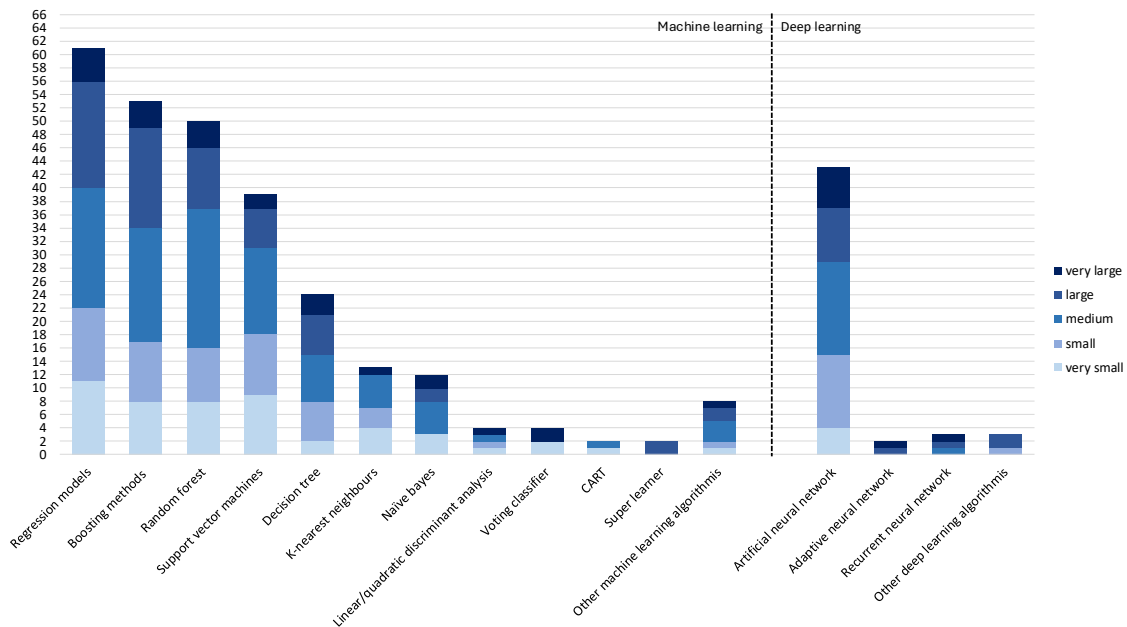


Figure 2.5: Frequency of algorithms used on datasets with very small (fewer than 300), small (300-700), medium (701-2000), large (2001-20000) and very large (more than 20000) sample size.

2.3.4 Model Evaluation

The evaluation metrics varied across the studies. Area under the curve (AUC) is most commonly used ($n=60$, 64%), and 32 (53%) of these papers used this value to assess model performance with imbalanced classes. Other frequently used performance metrics are recall ($n=44$, 47%), accuracy ($n=43$, 46%), and specificity ($n=34$, 36%). The majority of the studies used multiple performance metrics ($n=83$, 88%). Variable importance analysis is performed by 64 studies (68%) 61 of which (95%) reported PROMs data being valuable for prediction. Seventy-nine papers (84%) provided information on the best performing model. Regression models are the most frequently selected as best-performing algorithms ($n=24$), followed by boosting methods ($n=20$), random forest ($n=10$) and neural network ($n=10$).

Table 2.2: Machine learning and deep learning models used on data collected in one and multiple timepoints, ordered by number of publications.

Machine Learning (ML)	One Timepoint	Multiple Timepoints
Regression (n=61)	[91–101] [75, 102–111] [102, 112–121] [6, 122–131] [132–135]	[19, 136–144] [90, 145, 146]
Boosting (n=53)	[75, 93, 94, 98–102, 105, 147, 148] [106, 109, 110, 112–114, 149–152] [6, 102, 117, 119, 121, 123, 124, 153–155] [34, 95, 126, 129, 132, 133, 135, 156–158] [130, 134, 159]	[19, 90, 136–138, 140–142, 160, 161]
Random Forest (n=50)	[75, 91–95, 98, 103, 105, 147, 162] [106–109, 111–114, 117, 150, 163] [76, 119, 121, 122, 124, 125, 129, 135, 153, 158, 164] [6, 123, 131, 133, 134, 155, 159]	[88, 90, 136, 137, 139, 145, 146, 161, 165, 166]
Support Vector Machine (n=39)	[75, 91, 93, 94, 96, 99, 103, 104, 106, 107, 109] [88, 111–113, 117, 118, 150, 154, 167, 168] [76, 122–124, 126, 131, 134, 147, 164]	[19, 137, 140, 142, 145, 146, 160, 161]
Decision Tree (n=24)	[99, 101, 103, 118, 150–153, 167, 168] [6, 34, 76, 119, 134, 154, 156, 158, 159]	[19, 136, 138, 160, 165]
K-Nearest-Neighbours (n=13)	[96, 98, 99, 101, 109, 117, 122, 134, 169]	[88, 142, 145, 161]
Naïve Bayes (n=12)	[76, 98, 105, 111, 118, 130, 133, 134, 152]	[88, 161, 165]
Voting Classifier (n=4)	[99, 105]	[90, 141]
Discriminant Analysis (n=4)	[76, 96, 100, 152]	None reported
Classification and Regression Tree (n=2)	[96, 107]	None reported
Super Learner (n=2)	[123, 129]	None reported
Other ML Methods (n=8)	Wide and Deep [110] Stochastic Gradient Descent [122] Bagging [6] Bayesian Updating Algorithm [126] Graphical Gaussian Model [127] Multivariate Adaptive Regression Spline [103]	Hierarchical Gaussian Process [143] Autoregressive Integrated Moving Average [88]
Deep Learning (DL)	One Timepoint	Multiple Timepoints
Multilayer Perceptron (n=43)	[75, 91, 92, 97–99, 101–103, 106, 148] [109, 113–115, 117, 150, 152, 167, 168, 170] [102, 118, 120, 122, 124, 126, 128, 154, 164, 171] [132–135]	[19, 88, 90, 138, 142, 144, 161, 165, 172]
Recurrent Neural Network (RNN) (n=3)	None reported	Long-Short Term Memory [88, 89] RNN with Gated Recurrent Units [90]
Other DL Methods (n=4)	Adaptive Neural Network [98] Stacking Algorithm [118] Bayesian Network Model [173]	Adaptive Neural Network [89]

Table 2.3: Methods of addressing temporality in time-series data, with sample size and frequency of self-reports. Three papers which used time-series data did not report how the temporality is addressed [142, 165, 166], and are not included in this table.

Level of addressing temporality	Method of addressing temporality	Description of the method	Sample size	Frequency
Addressed by the model	Recurrent Neural Network (RNN)	Data transformed to 3D array and fed in the LSTM model [88] Events encoded by Adaptive Net, pooled by LSTM model [89] RNN with GRU, considering each treatment as timestep [90]	823 9,500 105,129	Weekly Irregular* Irregular*
Addressed in the features	Measured change	Change of measurement from baseline [19] Change in mean measurements from baseline [146] Mean daily change from the 24-h baseline period [145] Change in symptom severity from previous report [137]	245 31,700 116 34	Every 90 days Daily Twice a day Weekly
	Binary outcome	Variable indicated if a report is followed by exacerbation event [139] Occurrence of symptom in any day of a time window [143] Dichotomised score one week following the prediction date [161]	2,374 182,991 210	Daily Daily (3 days) Weekly
	Feature for each timeline	Score added as an input feature at every measurement [141] Created a timeline of best overall responses (BORs) [160]	83 31	Weekly Weekly
Not considered	Model for each timeline	Treated the 2- and 8-week measures as if assessed at baseline [140] Three models that used 7, 14, and 21 days as inputs [172]	1,003 20	3 time-points 3 time-points
	Selected 1 value for analysis	If patient had multiple follow-up events, the first is chosen [136] The assessment with the highest overall score is used [138] Score is updated at each assessment [144]	494 11,761 212,615	Irregular* Bi-weekly Irregular*

*Irregular measurements indicate that the reports are completed at any clinical event that occurred.

2.3.5 Clinical Relevance and Adoption

No studies reported that the developed methods had been applied in the clinical practice. Although we acknowledge, that in such multidisciplinary research clinicians are generally involved in the study design, only 3 papers (3%) explicitly mentioned how clinicians contributed to the model development. They helped select input variables [92, 167], or creating a testing set [158]. No papers mentioned patients involvement in the model development or any part of the study design. The majority of papers reported age ($n=63$, 67%) and gender ($n=60$, 64%) of study participants and only 24 (26%) studies reported ethnicity. Papers are classified into 3 different categories, inspired by a previously conducted scoping review [174]: internal validation (one source of data used for training and validation, including cross-validation or holdout sample for validation on unseen data from the same dataset), external validation (the model developed on one dataset and then tested/validated on a completely new (i.e. external) dataset) or deployment ("integrated into a prototype application, and evaluated for its feasibility in clinical workflows" [174]). Based on these definitions, 81 papers (86%) are in the internal validation stage, 10 (11%) completed external validation, and 3 papers (3%) are in the deployment stage.

2.4 Discussion

This scoping review aimed to identify AI methods used on PROMs data to predict patient outcomes. The analysis of 94 papers allowed the exploration of algorithms applied on complex patient-reported data and revealed the opportunities, challenges and best practice recommendations for AI medical research involving PROMs. The main findings suggest the unclear predictive value of PROMs, as well as inconsistencies in data pre-processing and model development design and reporting. Furthermore, there is a lack of stakeholder engagement during study design.

2.4.1 Study Characteristics

Due to fragmented data collection of PROMs, incorporating them into AI systems is very challenging [175]. Therefore, the majority of papers in this review have small sample size (Figure 2.3). In orthopedics settings PROMs have been increasingly collected as a part of routine care [176], which explains the large proportion of orthopedics papers with medium-to-large datasets. The large sample size is common in mental health papers, which might be explained by mental health diagnostic tools being often based on PROMs [177]. Contrary to orthopedics and mental health settings, PROMs collection in other healthcare domains is very limited. The respiratory datasets are mainly either very small or very large. The papers with very large sample size are predicting outcomes related to COVID-19 pandemic, where mobile applications collecting PROMs became more common [178]. Most of the studies analysed data collected specifically for research studies, rather than in clinical practice, which might introduce biases related to inclusion and exclusion criteria. The peak in the use of AI methods for all domains is between 2021 and 2022. This recent increase is compatible with a scoping review on AI in healthcare, where 71% of studies are published between 2020 and 2022 [179].

2.4.2 Data Pre-Processing

Missing data in AI research is an important aspect to investigate, as it can lead to various biases [180]. PROMs data collection poses challenges, as missing PROMs reports is common, both at random (e.g., participant not completing reports) or not at random (e.g., patients becoming too ill) [181]. Therefore, the inconsistencies in reporting data quality in the analysed studies are concerning. The justification of using data imputation techniques is also poor, whilst most commonly used techniques (complete case analysis and mean/median/mode imputations) can frequently cause bias [180]. Only a small number of papers used KNN-based imputation, which can reach the accuracy of complete data with a low performance difference [182]. Another inconsistency between the papers is caused by various methods for handling class imbalance in classification tasks. Papers which reported class imbalance often did not attempt to balance the data, which prevents models from appropriate learning from the training set. Furthermore, most of these papers used AUC as a performance metric, which require the balances setting to avoid bias [183]. The papers which reported balancing data have also done it inconsistently and without justification. Balancing data prior to train and test split can cause issues in model validation, as real data is never perfectly balanced. Therefore, it is important to use test set unaffected by sampling methods.

2.4.3 Model Development

The studies in this review reported model development process inconsistently, with majority of studies missing model hyperparameters reporting. According to Jha et al (2023)[184], it should be "the ethical obligation" to document all stages of model development that are essential for the reproducibility of results. The missingness in data are also handled and reported inconsistently, which is an important step for reproducibility as well. The lack of large PROMs datasets prevents applying deep learning methods, which can be extremely useful in capturing patterns in high-dimensional data or dependencies that other algorithms can't capture [185]. Only the simple "vanilla" neural network is applied more often than some of the basic ML models. Studies which collected data in multiple timepoints often did not address the temporality at all or analysed time-series data through data pre-processing strategies and conventional ML models. Only 3 papers used DL models that are appropriate for temporal processing. These are for example LSTM or GRU methods. Most papers in this review chose various ways to include temporal information through feature engineering, as described in Table 2.3. Nevertheless, DL models have been more successful in accurate predictions of patients' outcomes when applied to time-series data, as they are able to process more complex dependencies in high dimensionality and temporality of medical data [186].

2.4.4 Model Evaluation

Most studies used multiple evaluation metrics, which allows between-studies comparisons and in-depth analysis of model performance. Variable importance analysis is also commonly conducted, which supports the explainability of AI models [187]. Even though variable importance suggests that PROMs are important for analysis in most papers, these studies did not compare performances of models with

and without PROMs in input data. Currently there are inconsistencies between studies which found that PROMs improve model performance [6, 188], and studies which did not find PROMs to have as meaningful impact on model performance as objective measures [110, 163]. Furthermore, the papers used multiple models, which allowed them to select the best-performing one. The analysis of these showed that most common models are rarely selected as best models (e.g., random forest is selected as the best model only 20% of the times). However, voting classifier is used only 4 times, but selected as the best model 3 times. This suggests that further studies should perhaps pay more attention to models that are used less-frequently, which have a potential to perform better.

2.4.5 Clinical Relevance and Adoption

Limited reporting of clinician and patient engagement in the study development is of concern. This process is crucial for ensuring a patient-centered studies [189]. Involving stakeholders also can help build trust of the public to AI researchers, and as a result support the implementation of the studied tools in clinical practice [190]. Another issue arising from this review is the lack of external validation of the model performance in the studies, which is an essential step to potential clinical adoption. Assessing the model performance in a different setting may show different model performance which might suggest bias in the original study [174]. This review shows that new models keep being developed on original problems, without taking the studies further and exploring their potential in the real-world settings. Furthermore, the majority of papers did not provide any information on the ethnicity of the study participants. Ensuring diverse study population is an essential ethical consideration and lack of ethnic information can further contribute to deepening healthcare inequalities [184].

2.4.6 Recent Literature

The search of literature for this scoping review is completed in November 2023, and the paper is published in April 2025 [1]. Since the completion and publication of the initial scoping review some further relevant studies are published which are summarised here. As concluded above, it is very common to predict outcomes in the orthopedics domain from PROMs [191] using simple ML models in recent papers. For example, Lundgren et al., [192] predicted functional outcomes from pre-operative PROMs, clinical and demographic information, finding that LR is the best performing model with most important variables being preoperative PROMs, high BMI and femoral resection depth (posterior and distal), supporting functional positioning principles. Another study investigated the prediction of 90-Day readmission of total joint arthroplasty using EHR and PROMs [193]. As in the previous example, LR is the best performing model, but the most predictive variables are mainly clinical (comorbidities and treatment). There is also an example of a deep learning model predicting rehabilitation success, applying hybrid Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to time-series clinician and patient-reported data [194]. The DL model outperformed simpler ML models, and no explainability techniques have been attempted. In oncology, a paper predicting cancer-related fatigue in breast cancer survivors 15

years after diagnosis found that LR is the most successful model [195]. The three most important features are GP visits frequency before diagnosis of breast cancer, location of the tumour, and age at diagnosis.

As identified in the review, in the recent literature LR is one of the best performing models predicting outcomes from PROMs. The predictive value of PROMs is still unclear, even though explainability analysis are performed by most of papers. There is a variety of clinical and PROMs data selecting as most important features in the described studies. No direct comparison of models including or excluding PROMs is performed. The handling of missing data are reported in all the papers, which might indicate improvement in considering and reporting data quality in the field. Class imbalance is handled inconsistently. For instance, Park et al., [193] acknowledged the imbalance in data using appropriate performance metrics, but did not use any techniques to balance the data. Mahmoud et al., [194] used Synthetic Minority Over-sampling Technique (SMOTE) methods to oversample data. However, it is unclear at which stage the technique is used and if test set is affected by this method. Utilising time-series PROMs data are only performed by Mahmoud et al., [194]. The comparison of DL and ML provide an insight into the potential of more complex models to capture patterns in temporal PROMs. However, using only the CNN-RNN combination and simple ML models allows only the comparison of the high complexity of hybrid model with low ability to process high dimensionality in data by ML. Therefore, there should be more evidence of DL models, which require fewer parameters for analysing PROMs datasets, which are typically small. No reports of stakeholder engagement are found in the selected studies.

2.4.7 Strengths and Limitations

The main strength of this review is that it identifies AI models applied on self-reported data for predicting patient outcomes in all healthcare domains. In contrast, most existing reviews focus on a single specific domain, which limits the exploration of methods used. This is also the first known literature review with focus on the methodological approaches to use PROMs in AI prediction models. The use of 6 different databases from both health and computer science field helped reach many relevant papers, which might have been omitted by reviews using limited number of databases. This review also explored the rigour of model development and evaluation reporting. It focused on clinical adoption potential, from the perspective of patients and clinicians involvement and ethical consideration of participants' diversity.

The limitations of the literature review presented in this chapter include the possibility of omitting the studies published in languages other than English. Only published studies are considered, which might affect the conclusions and further deepen the publication bias. Since the review focuses on rigour in model development, evaluation, and wider stakeholder engagement, it is important to note, that the results are based on what is reported, and not what is done in the included papers.

2.4.8 Conclusions

The analysis of 94 papers in this scoping review revealed the increasing use of PROMs data in AI healthcare research. It provides an insight into the predictive potential of PROMs and utilising such data to improve patient outcomes. Nevertheless, this review also has revealed the inconsistencies in study design and evaluation. Four main research gaps have been found:

1. Unclear and inconsistent added value of using PROMs across the studies.
2. Lack of robust approach for data pre-processing.
3. Rare involvement of stakeholders (patients and clinicians) in the study design.
4. Limited evidence on the different types of AI models which can process time-series PROMs to predict patient outcomes.

These gaps have revealed the need for a comprehensive guideline on utilising and pre-processing PROMs data, as well as developing AI models to predict patient outcomes. The original research in this thesis will focus on addressing these gaps, applying models on PROMs data collected throughout chemotherapy and further explore the potential of both static and time-series patient-reported data for healthcare outcomes predictions.

Chapter 3

Research Methodology

3.1 Introduction

The literature review, reported in the previous chapter, provided a good understanding of existing AI methods applied on PROMs. However, it exposed research gaps in the literature, which this thesis aims to address. The scoping review revealed the lack of clarity on predictive value of PROMs, the inconsistencies in model development and evaluation, the limited range of methods and data types used, and the lack of stakeholder involvement in study design processes. Identifying these gaps highlighted the need for establishing methodological guidelines to utilise AI and PROMs to predict patient outcomes.

The research reported in this thesis uses a variety of AI models, types of PROMs, and different approaches to data pre-processing, model development and evaluation. Stakeholder engagement is also an important part of this research, which further guides the direction of the research and methodological choices. The results and conclusions from the studies reported in this thesis are used to derive a framework, which proposes a methodological approach for using different types of PROMs in AI models predicting patient outcomes.

This chapter presents the overall research methodology, which is designed to address the research gaps identified in the scoping review (Chapter 2), and propose the framework for using predictive AI models and PROMs. First, the research questions addressed in this thesis are listed and explained (Section 3.2). Further, the Clinical Context Section (3.3) provides the description of the main research concepts and the clinical trial, where the dataset used in this thesis comes from. The dataset is described in detail in Section 3.4. This chapter also provides a brief overview of all research phases and the summary of overall methodology (Section 3.5).

3.2 Research Questions

All research questions are formulated following the thorough literature scoping review (Chapter 2). This review revealed that there is an increasing number of research studies taking patients' views into account when building predictive models in healthcare, through the inclusion of PROMs as predictors. However,

it also identified methodological limitations of utilising PROMs in AI predictive models. The main research gap which is addressed in this thesis is the lack of comprehensive methodological framework for using AI and PROMs to predict patient outcomes. The need for this framework is identified through the following research gaps, discovered in the scoping review (Chapter 2).

The overarching aim of this thesis is to develop a comprehensive methodological framework for using AI and PROMs to predict patient outcomes.

This aim could support the responsible handling of PROMs data and AI model development, which could minimise bias and increase the clinical relevance of research. Four research questions are formulated based on the identified research gaps to guide the research, which are used to derive the framework:

- **RQ1: Added value of PROMs:** *What is the added value of using PROMs in AI models predicting patient outcomes?*

There is an increasing number of PROMs being collected and used in clinical practice and clinical trials. However, there is not enough evidence that these types of data can be used in AI predictive modelling. The literature showed that there are benefits of including PROMs as predictors in AI models [6, 196], but some studies did not find PROMs to have as meaningful impact on model performance as clinical or demographic data [110, 163]. The proposed framework for including PROMs in AI prediction models can be useful, as PROMs data hold a lot of important information on patients. However, when the predictive value of PROMs is known, the motivation for collection and inclusion of PROMs in AI models may increase. Therefore, this research question aims to assess if the addition of PROMs in AI medical research has the potential to improve predictions of patient outcomes.

- **RQ2: Design robustness:** *How do data pre-processing methods affect the results of AI models predicting patient outcomes?*

This research question explores if methodological inconsistencies for handling class imbalance identified in the scoping review (Chapter 2) can bias the results. Proposing approaches which can mitigate the bias arising from data pre-processing methods is important to help improve the robustness and interpretability of further studies using PROMs to predict patient outcomes.

- **RQ3: Stakeholder engagement:** *How can patients and clinicians be involved in the design of AI models predicting patient outcomes?*

Involving patients and clinicians in the design of research is important to improve its clinical relevance [37] and increase public trust in AI. However, the scoping review (Chapter 2) revealed that stakeholders are rarely involved, which is an important issue to address. Therefore, this thesis provides the first step in the exploration of how patients and clinicians could be involved. It should be noted that, due to lack of resources during this PhD

project, no thematic analyses are performed, and the recruitment is limited to groups that are working with research already without efforts to improve the representativeness of the sample. This thesis aims to raise the awareness of the importance of stakeholder engagement and encourage more advanced practices in future research.

- **RQ4: Data and model type:** *What are the most effective AI models for predicting patient outcomes using time-series PROMs?*

Using PROMs collected at one time-point and simple ML models for patient outcomes predictions is the most common approach, as identified in the scoping review. PROMs are often collected over-time, so using only one time-point for analysis may miss important information. Therefore, exploring the potential of longitudinal data and comparing the performance of ML and DL models could further support the understanding a broad range of methods, as it would benefit the comprehensiveness of the proposed framework.

3.3 Clinical Context

3.3.1 Chemotherapy Toxicity

Chemotherapy is a commonly used therapy for cancer, which can be used as a sole method of treatment or alongside surgery or radiotherapy [197]. Chemotherapy can lead to a range of symptoms and side effects that can negatively impact patients overall health, quality of life and functioning [198]. Serious adverse effects of chemotherapy affect 44.5% of patients receiving the treatment and can lead to initial or prolonged hospitalisation, permanent disability, life-threatening situations, or death [199]. The research reported in this thesis investigates the contribution of patient reported symptoms and treatment toxicity measured by PROMs in predicting hospital utilisation and changes in HRQoL:

1. Hospital utilisation - refers to the use of healthcare services (e.g., outpatient clinic visits, calls to healthcare staff, hospital admissions, emergency calls) for the purpose of preventing, curing and managing health problems, as well as maintaining health and well-being [200].
2. Health-Related Quality of Life (HRQoL) changes - A multidimensional construct in health care, that captures physical, emotional and social well-being, using validated questionnaires, representing a type of PROMs [48].

3.3.2 eRAPID Clinical Trial

PROMs have been increasingly used in oncology practice. For example, they can be utilised to screen for functional problems, monitor response to treatment or deterioration of a disease [201]. Evidence suggests that the systematic use of these measures improves communication and decision making between doctors and patients, which can lead to better patient satisfaction with care and quality of life. The information collected through PROMs could support shared exploration of treatment options through in-depth analyses of current health state and quality of

life expectations [65]. Furthermore, patients' level of independence, understanding of the health issue and responsibility for the management of the condition can increase after they start measuring and reporting symptoms [71].

The research reported in this thesis uses data collected solely from a clinical trial called Electronic patient self-Reporting of Adverse-events: Patient Information and aDvice (eRAPID). The trial is designed to test the efficacy of an intervention which involved self-reporting of symptom severity during chemotherapy, and generating automated severity-based advice to patients to support self-management or prompt hospital contact [70] (Figure 3.1).

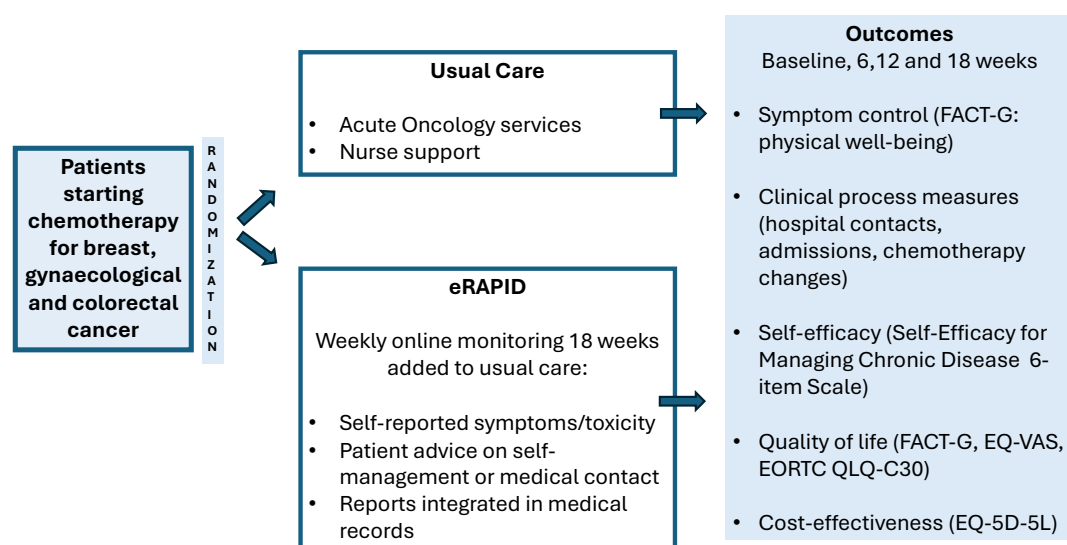


Figure 3.1: The eRAPID trial design. It involved 1:1 randomization (stratified by cancer site, gender, previous chemotherapy) to allocate patients to usual care or eRAPID intervention groups.

Trial Design

The eRAPID clinical trial was a prospective, randomized two-arm parallel group study. The participants were 508 patients initiating systemic treatment (chemotherapy with or without targeted therapies) for colorectal, breast, or gynecological cancers at the Leeds Cancer Centre (United Kingdom). The recruitment to the trial took place between January 22, 2015, and June 11, 2018. Patients were randomly assigned to usual care group (n=252) or intervention group (n=256). The usual care group was undergoing all standard healthcare procedures. Chemotherapy symptoms were assessed by nurses or clinicians regularly. The intervention group was asked to complete electronic symptom severity reports weekly in addition to receiving usual

care. The reminders were sent to patients via text. Following the report completion, the patients received immediate automated advice on self-management of symptoms, or were prompted to contact the hospital. The outcomes of the trial were symptom control, processes of care, patient self-efficacy, and global quality of life, measured with PROMs questionnaires at the baseline, 6, 12, and 18 weeks of chemotherapy. The trial design is presented in Figure 3.1.

Data collection

There are 2 different types of patient-reported data collected in this trial: the symptom severity reports used as an intervention, and the PROMs questionnaires used as outcome measures. Both are PROMs, but due to their different functions in the trial and collection process, only the outcomes of the trial are referred to as "PROMs", and the intervention data are referred to as "symptom severity reports". Therefore, these are the definitions of the data types collected by patients of the eRAPID clinical trial, which are used throughout this thesis:

- Symptom severity reports - Weekly reported electronic questionnaires, measuring the severity of symptoms experiences throughout chemotherapy. Each included multiple symptoms and a scale for assessing severity from 0 (no symptom) to 3 (severe). This questionnaire was created based on an established Common Terminology Criteria for Adverse Events (CTCAE)[59]. This is an instrument developed by National Cancer Institute, which can evaluate symptomatic toxicities by clinicians in cancer clinical trials. The grading scales and symptoms from CTCAE were transformed into patient-reported form for the purpose of the eRAPID trial. The examples of questions from the symptom severity reports, which patients were asked to complete weekly during the eRAPID clinical trial and the scale for severity measurements are provided in Table 3.1. The symptom severity reports were collected online and remotely.
- PROMs - Data collected through standardised, validated, and widely used questionnaires. In the eRAPID clinical trial they were used to measure the outcomes. Three different PROMs questionnaires were used:
 1. **The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ-C30).** This is a HRQoL questionnaire, consisting of 30 items, which contain information about participants' physical symptoms, perception on their physical function, emotional and social function and overall health and HRQoL [45]. This questionnaire is referred to in this thesis as QLQ-C30.
 2. **EuroQol 5 Dimensions 5 Level (EQ-5D-5L).** It included self-reported data on mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [202]. The visual analogue scale (health thermometer) involved rating of own health state on a scale 1-100. This questionnaire is referred to in this thesis as EuroQol 5 Dimensions 5 Level (EQ-5D)
 3. **Functional Assessment of Cancer Therapy-General (FACT-G).** This is a quality-of-life measure with 28 items, containing scores of Physical, Social, Emotional and Functional Well-being [46].

The schedule of collecting symptom severity reports and PROMs is presented in Figure 3.2.

Clinical and demographic data were also collected throughout the trial. The clinical data included the information on patients' health, such as comorbidities, cancer type, and previous treatment, collected through EHR. The demographic data was collected through a form at the clinical trial baseline.

Data on process of care were also collected throughout the trial. Information on acute admissions, patient-initiated calls to the hotline provided, and cancer treatment delivery (delays, dose modifications, and interruptions) were downloaded from electronic patient records [70].

Table 3.1: Examples of questions from the symptom severity reports from eRAPID clinical trial.

Symptom	Question	Scale
Pain	Have you had pain or discomfort anywhere on your body?	0: No 1: I had mild pain or discomfort 2: I had moderate pain or discomfort and I was not able to do some of the things I normally do (e.g. household chores, shopping) 3: I had severe pain or discomfort and I was not able to care for myself (e.g. getting out of bed, bathing, dressing)
Diarrhoea	Have you had diarrhoea (loose or watery stools)?	0: No 1: I have had diarrhoea and opened my bowels 2-3 times more in a 24 hour period than is normal for me 2: I have had diarrhoea and opened my bowels 4-6 times more in a 24 hour period than is normal for me 3: I have had diarrhoea and opened my bowels over 7 times more in a 24 hour period than is normal for me or I have been incontinent (unable to control my bowels)
Sleep	Have you had difficulty sleeping?	0: No 1: I occasionally have difficulty falling asleep, staying asleep or I wake too early 2: I often have difficulty falling asleep, staying asleep or I wake too early 3: I always have difficulty falling asleep, staying asleep or I wake too early

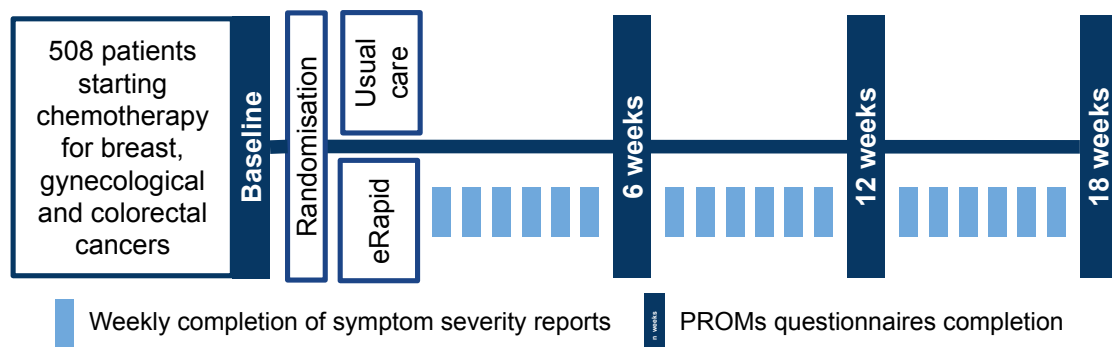


Figure 3.2: The expected schedule of collecting PROMs and symptom severity reports in the eRAPID clinical trial.

3.4 Dataset

The research reported in this thesis should be interpreted as a case study, because it presents the analyses performed on a single dataset. All research activities described in this thesis are conducted solely on the data collected in the eRAPID clinical trial. Therefore, all conclusions derived through this research provide methodological guidelines, rather than the evidence of performance of models predicting patient

outcomes during the chemotherapy. Ethics approval to use the data is granted by School of Medicine Research Ethics Committee [MREC 22-098], University of Leeds. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis adapted for machine learning research (TRIPOD+AI) checklist [203] is completed to self-assess the robustness of reporting and adapted to the whole thesis (Figures A3 and A4).

3.4.1 Predictor variables

The data used as predictors are PROMs, symptom severity reports, and clinical and demographic information, which are described in the previous section (3.3). They are used as input data in the predictive models developed throughout all research activities in this thesis. The number of all available predictor variables and information they hold are provided in table 3.2.

3.4.2 Target Variables

There are two types of outcomes predicted by AI models in this thesis: hospital utilisation and HRQoL changes during chemotherapy:

- **Hospital utilisation.** There are three different variables considered as hospital utilisation: hospital admissions (if the person is likely to be admitted to the hospital during chemotherapy), clinical triage (if the person is likely to contact the provided hotline number), and changes to chemotherapy (if a patient is likely to have any change of a drug or a dose throughout the treatment). When the event happened, it was recoded in electronic patient records, from where the data on these events were downloaded. Binary likelihood of these outcomes is predicted by developed models. To achieve this, the target variables are transformed to binary variables, where class 0 indicates no event, and class 1 indicates at least one occurrence of the event of interest.
- **HRQoL changes.** Five different HRQoL variables are considered: physical well-being, social functioning, role functioning, usual activities, and mobility. The data on these aspects were collected through standardised PROMs and used as predictor variables, as well as to calculate changes for prediction. In this thesis changes in these variables from baseline to 6, 12, and 18 weeks are predicted. The changes are measured by the difference in the reported PROMs between the time-points of interest. This difference is calculated using the MCID of a questionnaire which is used to measure a HRQoL aspect. The changes in these variables are predicted as a binary likelihood of improvement (1 indicating improvement of the outcome, 0 indicating no change or deterioration of the outcome) and deterioration (1 indicating deterioration of the outcome, 0 indicating no change or improvement of the outcome).

3.4.3 Data Management

Data Provenance

The dataset for this project was collected and provided by the eRAPID researchers. The symptom reports were completed entirely online by the intervention group. All

Table 3.2: Data types in the dataset used in the research reported in this thesis.

Variable type	Number of variables	Information	Collection type	Collection time-point
Predictor variables				
Demographic	6	age at study entry, sex, BMI, marital status, employment status, education level	Self-reported questionnaire	Baseline only
Cancer characteristics (clinical)	3	disease site (breast/gynecological/colorectal), previous chemotherapy (yes/no), information if the disease is metastatic or non-metastatic	Participants' medical notes	Baseline only
Comorbidities (clinical)	32	cardiovascular, respiratory, gastrointestinal, stomach/intestine, endocrine, renal, neurologic, rheumatologic, previous malignancy, and substance abuse	Participants' medical notes	Baseline only
QLQ-C30 (PROMs)	15	physical symptoms, physical, emotional and social function, overall health	Self-reported QLQ-C30 questionnaire (scores calculated for 15 aspects of HRQoL)	Baseline and at 6, 12 and 18 weeks
EQ-5D (PROMs)	5	mobility, self-care, usual activities, pain/discomfort, and anxiety/depression	Self-reported EQ-5D questionnaire	Baseline and at 6, 12 and 18 weeks
FACT-G (PROMs)	4	physical, social, emotional, and functional well-being	Self-reported FACT-G questionnaire	Baseline and at 6, 12 and 18 weeks
Symptom severity reports	21	pain, nausea, vomiting, diarrhoea, temperature, chills, physical ability, appetite, fatigue, sore mouth, indigestion, shortness of breath, reaction of injection site, phlebitis, sore hands, neuropathy, taste, nail changes, nose bleeds, skin rash, and ascites	Self reported symptom severity scores	Weekly
Target variables				
Hospital Utilisation	3	hospital admissions, triage events, changes to chemotherapy	Electronic patient records	When the event occurs
HRQoL Changes	5	physical well-being, social functioning, role functioning, usual activities, mobility	Changes in self reported PROMs	At 6, 12 and 18 weeks

PROMs were collected on paper, and required transforming to digital format. The dataset was provided for the research conducted for this thesis:

- **File 1:** Baseline information (clinical and demographic) and standardised PROMs (EQ-5D, QLQ-C-30 and FACT-G) scores obtained at the baseline, 6, 12 and 18 weeks. Rows were participants' IDs, and columns were clinical, demographic and PROMs data values.
- **File 2:** Symptom severity report scores. Rows were the time-step of each report and columns were the severity scores of the symptoms. Participant's ID is provided for each time-step.
- **File 3:** Hospital admissions that happened throughout the trial. Rows were the the time-steps of the events. The columns provided participant's ID and additional information about the event. Only the time-step information and participant's IDs were used in this thesis.
- **File 4:** Triage (contacting the provided hotline number) that happened throughout the trial. Rows were the the time-steps of the events. The columns provided participant's ID and additional information about the event. Only the time-step information and participant's IDs were used in this thesis.
- **File 5:** Changes to chemotherapy that happened throughout the trial. Rows were the the time-steps of the events. The columns provided participant's ID and additional information about the event. Only the time-step information and participant's IDs were used in this thesis.

All datasets were provided in a tabular form in SPSS, which were then transformed to Excel spreadsheets. All data provided was unprocessed from the point of data collection, apart from QLQ-C30 questionnaire (the scores were aggregated by eRAPID researchers for each domain of the questionnaire). The data selected for analysis differed between chapters. For Chapter 4, only data collected at baseline were selected for analysis. These included PROMs, clinical and demographic information obtained from File 1. The files with hospital utilisation information (3-5) were not merged with the data, as only total numbers of admissions, triage events and changes to chemotherapy are calculated. These numbers were then transformed to binary outcome for prediction (indicating no event or at least one event of hospital utilisation). File 1 was also the only one used for the analysis in chapter 5, as it included data on PROMs collected at different time-points of clinical trial. The differences from baseline values were calculated and transformed to binary outcomes of improvement and deterioration. Chapters 6 and 7 used all available files. File 2 provided the structure of the dataset. For chapter 6, the information from file 1 was repeated for each time-step of a symptom report completed by the same participant. For chapter 7, file 1 was kept separate, as the model used in this chapter enabled separate inputs of different structures. For both chapters, the time-steps of admissions and triages from File 4 were used to calculate two binary variables, indicating if the event happened within 14 days from the symptom report, and if the event happened between the time of report completion and the end of the trial.

Data Quality

The inevitable data quality issue of longitudinal patient-reported data is the irregularity of reporting. The participants of eRAPID clinical trial were asked to complete the symptom severity reports weekly [70]. The irregularity of reporting varied between patients, which resulted in many missed or too frequent symptom severity reports. The frequency of reporting per participant is plotted in Figure 3.3. It shows that out of 256 participants in the intervention group of the trial, fewer than 50 provided the expected 18 reports. Most participants provided fewer reports than 18. There were also people who reported more often than asked. This irregularity introduces challenges in data analysis, as it can provide hidden information on patients' health status. For example, missed reports can indicate severe symptoms which might have impacted the ability of patients' to fill in the report. Awareness of these challenges and discussing them with patient representatives led to adding a variable indicating time from previous report. This enabled the use of this dataset for the prediction of chemotherapy outcomes, with the consideration of irregularity in data.

The predictor variables used as outcomes in this thesis are often imbalanced. For example, in the available dataset most participants were not admitted to the hospital than admitted. ML models perform worse when they are trained on imbalanced data and the results might be skewed towards the majority class (no admission). Addressing irregularity in data is a common challenge in medical AI research, but often handled without consideration on how the balancing methods can bias results [12]. Due to inconsistent methods for addressing class imbalance in literature, this thesis aims to compare different balancing methods and explore their impact on model performance. the comparison provides important insights

into the bias which data pre-processing methods can cause.

Another challenge of this dataset is that the data were collected as a part of an intervention. The eRAPID clinical trial aimed to explore the effectiveness of online monitoring and immediate patient advice on the management of control better treatment- and disease-related symptoms during chemotherapy. One of the automated advice that patients in the intervention group were receiving was a hospital contact, which could link to the outcomes of hospital utilisation. To assess the risk of bias resulting in study arm allocation, models with and without the "study arm" variable, random forest (RF) models including all variables predicting admissions, triage and chemotherapy changes were run and tested on a) only control group data; b) only intervention group data. The largest difference between the values of AUC was 0.05. Based on these results and the fact that the clinical trial also did not find the intervention group to be different from the control group for most of the outcomes, the potential bias of the intervention group can be rejected. The paper reporting eRAPID clinical trial also found that the hospital utilisation did not differ significantly between the intervention group and the usual care group.

3.5 Methodology Overview

An overview of the research methodology applied in this thesis is presented in Figure 3.4. A range of methods were utilised including evidence synthesis and ML and DL model development with ongoing stakeholder involvement. The figure provides brief descriptions of the aims and outcomes of each phase of the research. The phases are described below in more detail.

3.5.1 Description of Research Phases

Each chapter of this thesis presents a different research phase. Chapter 2 identified gaps in literature, which helped formulate research questions. Each of the research questions stated in the Section 3.2 of this chapter are addressed gradually by Chapters 4, 5, 6, and 7. Each question appears in multiple chapters to ensure comprehensive and in-depth analysis and discussion on every aspect that this thesis is addressing. Table 3.3 presents the summary of all stakeholders involved in the design of research reported in this thesis.

Literature scoping review (Chapter 2):

The first phase of this research is a literature review, which explores the research gaps to formulate research questions. There are three main types of literature reviews: systematic, scoping, and narrative. Systematic reviews focus on a specific and narrow research question, which does not capture the breadth of methods well. Narrative review, even though increasingly common, does not involve systematic search of literature, which can lead to bias [204]. Scoping review has been chosen as a type of the literature review, as it enables a broad exploration of AI methods implemented in existing research to analyse PROMs [205]. The review involved systematic search of six databases and the analysis of 94 selected papers

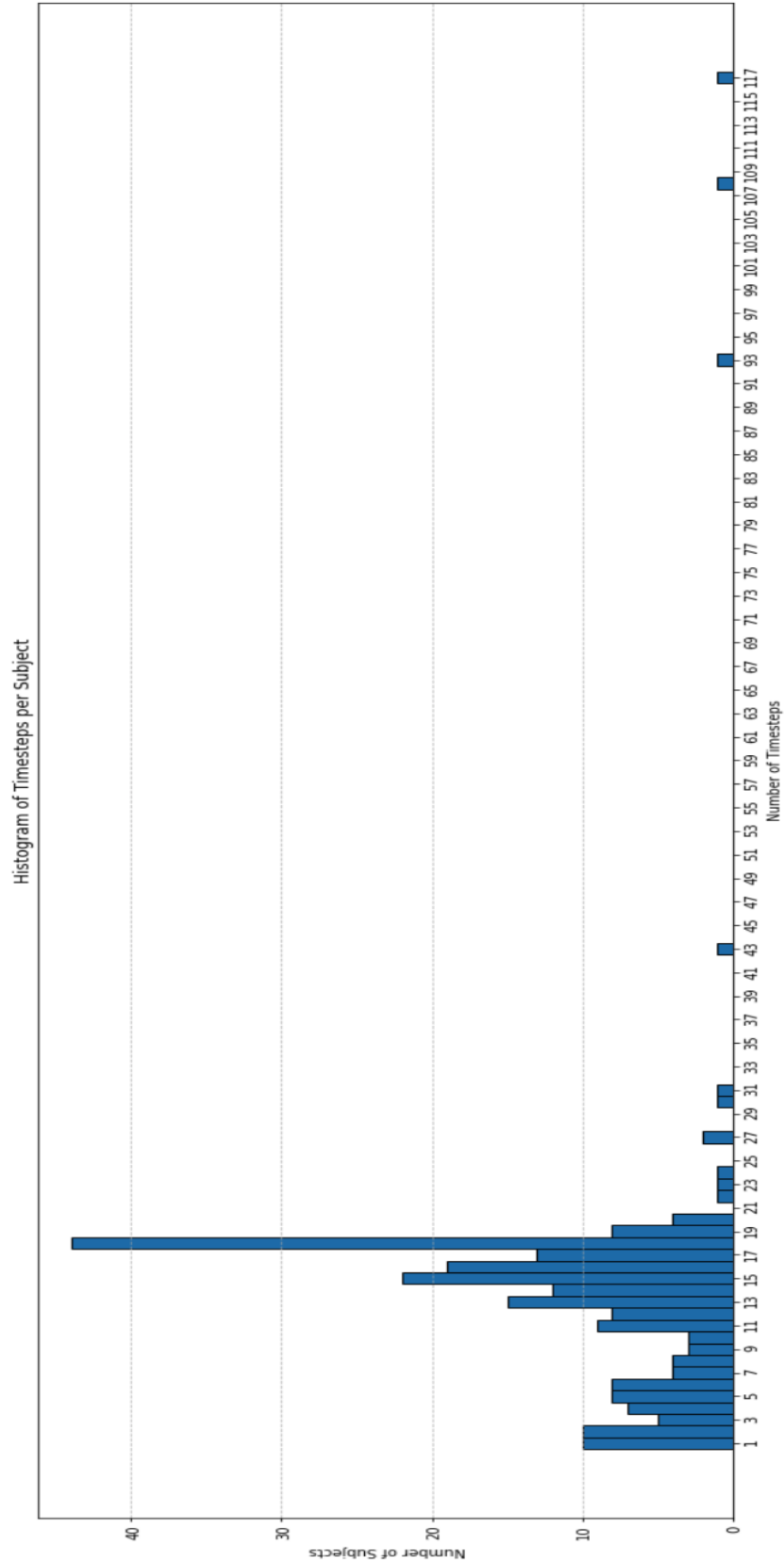


Figure 3.3: A frequency plot of the number of time-steps (reports) completed.

Literature scoping review Aim: To identify and summarise AI methods applied to PROMs for prediction of patient outcomes. Input: Data extracted from research papers selected from 6 different databases. Output: Research gaps which highlighted the need for methodological guidelines for using AI and PROMs to predict patient outcomes. Identified Gaps: 1. Added value of PROMs: Unclear and inconsistent added value of using PROMs across the studies. 2. Design robustness: Lack of robust approach for data pre-processing. 3. Stakeholder engagement: Rare involvement of stakeholders (patients and clinicians) in the study design. 4. Data and model type: Limited evidence on the types of AI models which can process time-series PROMs to predict patient outcomes. The identified gaps informed the design of this research		Chapter 2 Health and Quality of Life Outcomes Journal			
Predicting hospital utilisation from baseline PROMs Addressed gaps: 1, 2, 3 Input: Baseline PROMs, clinical and demographic data. Output: Binary prediction of hospital admissions, triage events and changes to chemotherapy.		Added value of PROMs Explored through comparing model performances including and excluding PROMs, and feature importance.	Design robustness Identified bias in inappropriate irregularly handling technique. Grid search for hyperparameter tuning. Evaluation with multiple performance metrics.	Stakeholder engagement Clinician involved with data pre-processing, variable selection and outcome selection.	Data and model type Baseline PROMs collected at 1 time-point analysed with classical ML.
Predicting quality of life changes from baseline PROMs Addressed gaps: 1, 3 Input: Baseline PROMs, clinical and demographic data. Output: Binary prediction of deterioration and improvement in physical well-being, social functioning, role functioning, usual activities, and mobility at 6, 12 and 18 weeks from starting chemotherapy.		Explored through feature importance only.	Test set unaffected by irregularity handling. Grid search for hyperparameter tuning. Evaluation with multiple performance metrics.	Patients and clinician involved with outcome selection. Clinician involved with data pre-processing, variable selection.	Baseline PROMs collected at 1 time-point analysed with classical ML.
ML and time-series PROMs for hospital utilisation and HRQoL changes prediction Addressed gaps: 1,2,3,4 Input: Longitudinally collected symptom severity reports, PROMs, baseline clinical and demographic data. Output: Binary prediction of admission happening within 14 days from a symptom severity report and admission at any time during chemotherapy; Binary prediction of QoL changes.		Explored through comparing model performances including and excluding PROMs, and feature importance.	Further prevention of bias and data leakage with data pre-processing completed within nested cross-validation for hyperparameter tuning. Evaluation with multiple performance metrics.	Clinician involved with data pre-processing, variable selection. Outcome follows previous chapters.	Longitudinal PROMs collected at many time-points analysed with classical ML.
LSTM and time-series PROMs for hospital utilisation and HRQoL changes prediction Addressed gaps: 3, 4 Input: Longitudinally collected symptom severity reports, PROMs, baseline clinical and demographic data. Output: Binary prediction of admission and triage event happening within 14 days from a symptom severity report and admission and triage event at any time during chemotherapy.		Explored through feature importance only.	As in Chapter 6, prevention of bias and data leakage with data pre-processing completed within nested cross-validation for hyperparameter tuning. Evaluation with multiple performance metrics.	Patients and clinician involved with data pre-processing, variable selection and outcome selection.	Longitudinal PROMs collected at many time-points analysed with classical ML and DL.
Chapter 7 IEEE ICHI Conference		Chapter 6 MIE Conference			

Figure 3.4: Overview of Research Methodology.

Table 3.3: Summary of stakeholders involved in the the design of research reported in this thesis.

Type of stakeholders	Stakeholders	Description	Consulted chapters	Type of consultation	Date of consultation
Clinical	Professor Galina Velikova	Consultant in Medical Oncology and a Professor at the University of Leeds. Involved in this research as a supervisor.	All chapters	Continuous involvement as a project supervisor	Continuously throughout the project
	Patient Centred Outcomes Research (PCOR) group	A group of clinical oncologists, nurses, healthcare scientists and psychologists who work on applied healthcare research aimed at improving cancer patients' quality of life. The group is based at the St James's Institute of Oncology, University of Leeds.	Chapter 4	Presentation followed by discussion	April 2023
Patient representatives	Three patient representatives from "Use MY data"	"use MY data" organisation, which is an independent movement of patients, relatives and carers focused on the use of patient data in research.	Chapter 5	Poster presentation and discussion at the Cancer Research UK conference	February 2024
	Three patient representatives from the Research Advisory Group	Patient representatives recruited from the research advisory group working with PCOR group at the University of Leeds	Chapters 6 and 7	Individual interviews	December 2024

Website of the PCOR group: <https://pcor.org.uk/>. Website of the "use MY data": <https://usemydata.org/>.

which used PROMs and AI methods to predict patient outcomes. It summarised AI methods used on PROMs across various healthcare domains and identified the trends and research gaps. It focused on exploring:

- Study characteristics - to provide the overview of most frequently used data types and to understand the trends in research overtime.
- Data pre-processing methods - to explore how missingness, irregularity and class imbalance were handled before analysing the data.
- Model development - to explore the model selection and techniques for optimising hyperparameters.
- Model evaluation - to understand how the performance of models is assessed.
- Clinical relevance and adoption - to investigate if stakeholders were involved in study design, and to understand if adoption potential of the research is explored.

The analysis of these aspects revealed inconsistencies in study design. Understanding the research gaps helped formulate research questions stated in Section 3.2 of this chapter.

Predicting hospital utilisation from baseline PROMs (Chapter 4):

All analyses performed in this chapter use PROMs, clinical, and demographic information (Table 3.2) collected at the baseline of the eRAPID clinical trial. The temporality of data is not considered in this chapter, as focusing only on data collected at one time-point can provide more insights into the inconsistencies in methods used in previous studies (revealed in the scoping review), where using data from single time-point is the most common approach. This chapter predicts the binary likelihood of hospital admission, triage (contact made with hospital) and change to chemotherapy throughout the treatment. Six different ML models are used as classifiers in this chapter: Logistic Regression (LR), Decision Tree (DT), Adaptive Boosting (AB), Random Forest (RF), Support Vector Machines (SVM), and Multilayer Perceptron (MLP). These were selected due to their frequent use

in similar research and the variety of classification types, explained in detail in Chapter 4. The chapter aims to begin addressing the research questions exploring:

- **Added value of PROMs** - through comparing the performance of models applied on data including and excluding PROMs. Furthermore, variable importance analyses are performed to explore which information is the most relevant for models when making predictions.
- **Design robustness** - through comparing the performance of models developed using data affected by three different ways of handling class imbalance. As reported in the previous chapter, Section 2.2.4, class imbalance can affect model performance, and inconsistent use of these methods in literature affects the interpretation and between-studies comparison of results. There are three different ways of addressing class imbalance in papers applying ML models on PROMs in literature:
 1. Balancing dataset prior to splitting the data to train and test set.
 2. Splitting the data to train and test sets, then balancing train set only.
 3. No techniques for data balancing used.

Comparison of the performance of models built on datasets affected by each of these methods can help understand the bias caused by data pre-processing techniques.

- **Stakeholder engagement** - through the consultation of study design with a clinical oncologist and the Patient Centred Outcomes Research (PCOR) Group in the University of Leeds, Faculty of Medicine and Health, Leeds Institute of Medical Research. This group includes health scientists and clinicians. The preliminary methods and results were presented to PCOR group and the feedback was taken into consideration when designing the research reported in Chapter 4.

Predicting HRQoL from baseline PROMs (Chapter 5):

Chapter 5 further utilises PROMs completed at the baseline of the eRAPID clinical trial (Table 3.2). The predicted outcomes are the changes in HRQoL throughout chemotherapy. These outcomes were selected as they were considered important by patient representatives consulted during the design of the research reported in Chapter 5. Models used for predictions were reviewed from the Chapter 4, and in this chapter they are: LR, Extreme Gradient Boosting (XGBoost), RF, and MLP. XGBoost is a preferred boosting method over AB, as it is known to have the potential to outperform other ML models on tabular data [206]. The data pre-processing methods follow the recommendations from Chapter 4, so only a train set is balanced, leaving test set imbalanced. Chapter 5 explores the answers to the following research questions:

- **Added value of PROMs** - through showing the potential of using PROMs to understand how different aspects of HRQoL can change at different stages of chemotherapy. In this chapter, baseline PROMs are used as input variables, and time-series PROMs are used to calculate a change in HRQoL from baseline

to different time-points during treatment (6, 12, and 18 weeks of treatment). The changes in physical well-being, social functioning, role functioning, usual activities are predicted for all time-points. A clinical oncologist is also involved throughout the design of research in this chapter.

- **Stakeholder engagement** - through a consultation with patient representatives. During the study design, patients representatives provided the information on the meaningful outcome of the ML prediction models. They mentioned that it would be very beneficial for them to expect how chemotherapy would impact HRQoL, rather than only clinical aspects (such as hospital utilisation explored in Chapter 4).

ML and time-series PROMs for hospital utilisation and HRQoL changes predictions (Chapter 6):

The previous chapters explored the ML applied to baseline patient-reported, clinical, and demographic data to predict outcomes meaningful for both clinicians and patients. However, the longitudinal data remained unused. Therefore, this phase is the first one in this thesis to develop ML models applied on time-series weekly symptom severity reports, PROMs collected at 3 time-points, and clinical and demographic data (Table 3.2) to predict both hospital utilisation and HRQoL changes during chemotherapy. Since ML models were most frequently used for similar tasks in the literature, 4 ML models (the same ones as in Chapter 5) are applied to longitudinally collected data. The outcomes of admission and triage happening at any time between when the report was completed and the end of study time-point at 18 weeks (long-term predictions) and within 14 days from the completed report (short-term predictions) are considered to understand the models' ability to identify short-term and long-term patterns in time-series data.

This chapter addresses all four research questions:

- **Added value of PROMs** - through comparing the performance of models including and excluding PROMs where time-series symptom severity reports, demographic and clinical data are available for the models to learn from. This comparison aims to establish if standardised PROMs questionnaires are useful for predictions, even when weekly symptom severity reports are available. Variable importance analyses are also performed to understand the predictive potential of all variables.
- **Design robustness** - through building on the recommendations from the previous chapters and adding design components which have the potential to further mitigate bias. In AI modelling, data leakage is an accidental sharing of information between train and test sets which can result in performance estimates being better than in the intended application scenario [13]. Splitting data to train and test set prior to balancing in Chapters 4 and 5 was mitigating data leakage, but overlooking the risk arising from using missing data imputation methods on prepared dataset. This chapter uses a nested Cross-Validation (CV) method, which enables all data pre-processing steps to be performed in the individual loops of the CV design, preventing data leakage.

This method is also considered to more robustly optimise hyperparameters [207].

- **Data and model type** - through exploration of the performance of ML models applied on data collected at multiple time-points. These include weekly symptom reports, PROMs questionnaires completed at the baseline, 6 and 12 of the chemotherapy and clinical and demographic data, which are static (collected at one time-point - baseline). This chapter only begins exploring this research question, as it uses only simple ML models with limited potential to process high-dimensional data. However, investigating the ability of these models to perform short- and long-term predictions of outcomes can inform if more complex methods are necessary to use.
- **Stakeholder engagement** - through the consultation of study design with a clinical oncologist and patient representatives. The selection of clinically relevant time-series variables followed the discussion with a clinical oncologist. The involvement of patient representatives enabled understanding of patterns of reporting and reasons for irregularity, which supported data pre-processing for this chapter.

DL and time-series PROMs for hospital utilisation and HRQoL changes predictions (Chapter 7):

The main aim of this chapter is the comparison of performances of the ML models developed in Chapter 6 with a Long-Short Term Memory (LSTM) model. LSTM is an example of a DL model, which is known for its ability to process time-series data well [208]. This model was selected for analysing the data in this chapter, as it was the most frequently used model in the limited examples of processing temporal PROMs, as reported in Chapter 2. Nevertheless, there is limited evidence of the differences in performance of ML models and LSTM applied on time-series PROMs combined with static clinical and demographic data for both short-term and long-term predictions of outcomes. This chapter further addresses the following research questions:

- **Data and model type** - through a critical comparison of LSTM model and ML models for long- and short-term predictions of hospital utilisation and prediction of changes in HRQoL.
- **Stakeholder engagement** - Through discussion on irregularity of reporting and using DL methods in medical research. The irregularities in reporting form additional challenge for LSTM, as this model requires equal number of reports for each patient to process time-series data. There are two main ways of applying LSTM model on irregular data:
 1. Imputing missing reports and aggregating the ones which were completed too frequently.
 2. Padding sequences to match the longest sequence (using a value which does not exist in the dataset to artificially fill in sequences) [194]. A masking method is then used to inform an algorithm to ignore the padded values when predicting outcomes, and only use original data.

To decide between the two methods, patient representatives were consulted to understand what are likely reasons for skipping reports or for reporting too frequently. After the consultation it is clear that the reasons for irregularity can be informative for the model to learn from. Therefore, the first option to handle irregularity (aggregating and imputing reports) is not considered, as it would hide the irregularity from the dataset. Instead, padding sequenced is performed to ensure equal sequences required by LSTM models.

3.5.2 Summary

The methodology presented in this chapter included the consideration of four main research gaps identified when reviewing the literature reporting research using AI and PROMs to predict patient outcomes (Chapter 2). The research questions, formulated based on the research gaps, guided the exploration of the best approaches for applying AI methods on PROMs data. Figure 3.5 presents the summary of the methodology, including the information on research gaps identified in the scoping review and which research questions are addressed by which chapters of this thesis.

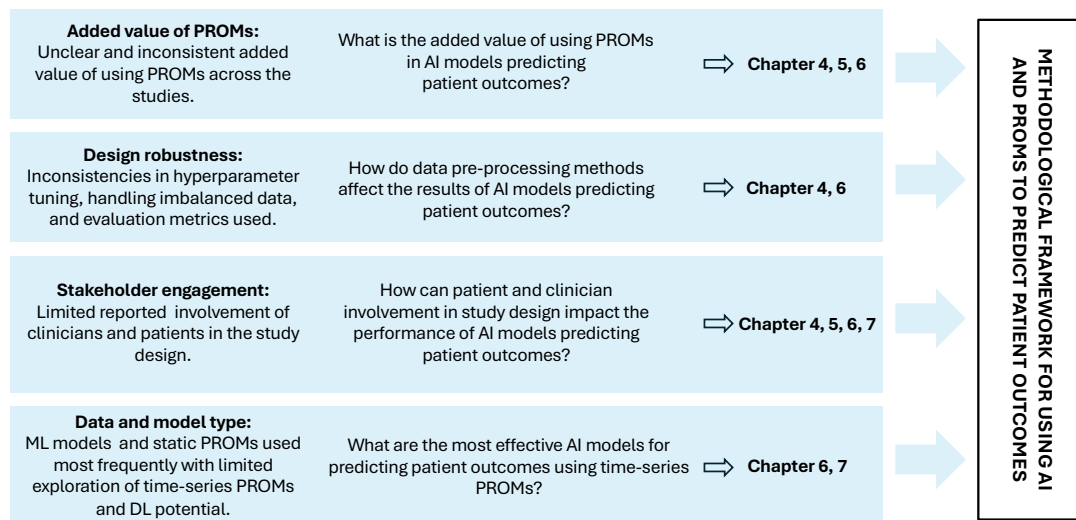


Figure 3.5: Summary of the research methodology as it is developed in the thesis. Research questions were formulated to address the limitations of data and model type, design robustness, and stakeholder engagement. The chapters where the corresponding questions are addressed are shown. Time-series PROMs refer to both symptom severity reports and standardised questionnaires collected longitudinally.

Chapter 4

Using Baseline PROMs To Predict Hospital Utilisation

4.1 Introduction

The scoping review (Chapter 2) explored AI methods which have been applied on PROMs to predict patient outcomes in previous studies. It has shown that there is an increasing number of research studies taking patients' views into account when building models in healthcare using AI and PROMs to predict patient outcomes. The review has also highlighted limitations in explored studies, such as inconsistencies in data pre-processing methods and model development, unclear predictive value of PROMs and limited stakeholder involvement in research design. As described in Chapter 2, PROMs used as input data in models predicting patient outcomes were usually collected at one time-point. However, data handling techniques differed between studies. Therefore, this chapter, using PROMs completed at the one time-point before chemotherapy, begins addressing the research gaps identified in the scoping review (Chapter 2) and exploring answers to the following research questions:

1. **What is the added value of using PROMs in AI models predicting patient outcomes?** This is the fundamental research question to address in order to assess the need for the framework of applying AI models on PROMs for the prediction of patient outcomes. As shown in Chapter 2, there is an inconsistent evidence on predictive value of PROMs in literature. Therefore, this chapter compares the performance of models trained on data which include and exclude PROMs. Variable importance analysis further helps understanding the predictive value of PROMs.
2. **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** Specifically, this chapter explores how different methods for handling class imbalance impact results of AI models. It compares three different approaches in order to understand how they influence the risk of bias.
3. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** The research reported in this chapter was designed with the involvement of a PCOR group and a clinical oncologist (Chapter 3, Table 3.3). The stakeholder engagement helps understanding

which aspects of the design can be supported by involving clinicians. Patient representatives have not been consulted at this stage, so their involvement will be explored in further chapters.

The research reported in this chapter uses data collected in the eRAPID clinical trial, described in Chapter 3, Section 3.4. It focuses on predicting hospital utilisation outcomes during chemotherapy. Chemotherapy toxicity is associated with physical, psychological, and psychosocial symptoms negatively impacting patients' HRQoL. Adverse effects of chemotherapy can lead to initial or prolonged hospitalisation, permanent disability, life-threatening situations, or death [199]. Unplanned healthcare utilisation (including hospital admissions, triage events and changes to treatment) due to chemotherapy complications can be detrimental for the healthcare quality and patients' physical and emotional well-being, as well as increasing burden on healthcare system [209]. Early identification of chemotherapy toxicity symptoms and close monitoring of patients are essential to enable prevention of serious long-term health issues and mortality [210]. Therefore, understanding factors that contribute to acute presentations to hospitals in cancer patients undergoing treatment can support planning for emergency admissions, and as a result, increase the quality of care and reduce healthcare costs [211]. Identifying and characterising patients at risk of chemotherapy-related healthcare utilisation could support the healthcare system through informed planning for hospital presentations and admissions. Furthermore, understanding factors that might be related with increased chemotherapy toxicity can potentially personalise cancer care through enabling an informed shared decision-making process when choosing cancer treatment [212, 213].

ML methods have been used to successfully achieve prediction of patient outcomes using various clinical data. However, PROMs are often not included in the analysis, even though these measures hold important information about patients' perspective on their health status. The literature review indicated that including PROMs in healthcare prediction models could help identify patients at risk of experiencing undesirable clinical outcomes [196]. For instance, ML algorithms trained on patient-reported and clinical data accurately predicted financial toxicity (negative economical burden of treatment [214]) in early breast cancer patients [188]. Furthermore, PROMs enhanced ML algorithms performances predicting 5-year cancer survival, when added to clinical and sociodemographic variables [6]. Nevertheless, some studies did not find PROMs to have meaningful impact on model performance [110, 163].

To the best of our knowledge, the predictive value of PROMs is not explored in detail because of the variety of PROMs instruments currently used in literature [73], inconsistent performance metrics which affect between-study comparisons of results, and the lack of clarity on how inappropriate data pre-processing methods can bias the results. For example, balancing datasets before creating train and test sets, which often introduces bias [12], but it is still a common practice in literature. The lack of methodological agreement and guidance indicates the need for comparison of frequently used methods. PROMs completed at one time-point were the most frequently used in studies reported in Chapter 2, which suggests their non-complex structure. However, the methodological inconsistencies were present even for these simple PROMs datasets. Therefore, this chapter assesses if PROMs

completed at baseline of chemotherapy improve performance of machine learning (ML) models predicting hospital admissions, triage events (contacting helpline or attending hospital), and changes to chemotherapy throughout the treatment. It also explores how each of the available PROMs questionnaires impacts the prediction of outcomes individually. Finally, this chapter investigates how three different methods for handling class imbalance can affect model performance.

4.2 Methods

4.2.1 Methodology for Predicting Hospital Utilisation from Baseline PROMs

This chapter presents the application of six different ML models to predict three target variables related to hospital utilisation: hospital admissions, triage events and changes to chemotherapy. They were selected due to the negative impact of unplanned hospital utilisation on clinicians and patient, as understood through the consultation with a clinical oncologist and through studying relevant research papers [209, 212, 213]. The models were applied on six different variable sets, which involved PROMs, clinical and demographic data in different configurations to compare the predictive value of PROMs. Furthermore, three different pre-processing methods for class imbalance in data were used in order to understand how they affect the results. The overview of the methods is presented in Figure 4.1.

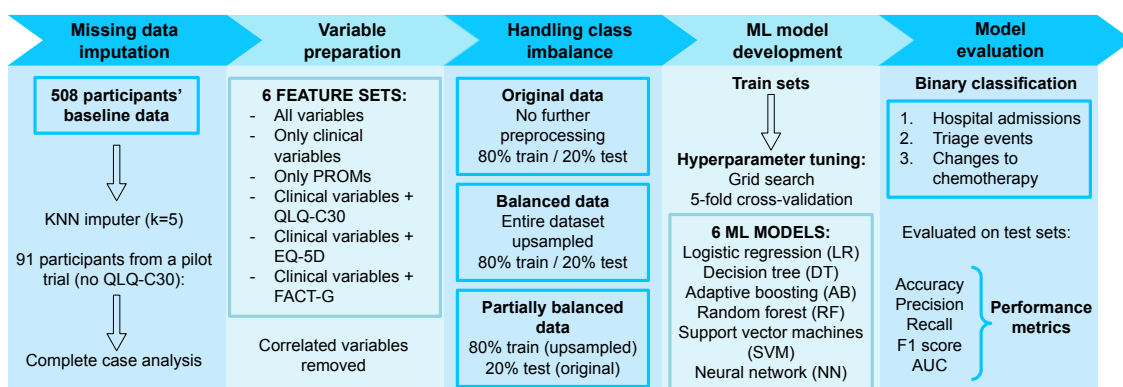


Figure 4.1: Flow diagram illustrating the methodology of the study.

4.2.2 Dataset

Input Variables

Data from 508 patients were collected in the eRAPID clinical trial [70], described in Chapter 3. All variables selected for this study were chosen based on availability and clinical relevance discussed with a clinical oncologist. Out of the entire dataset (see section 3.4, Table 3.2), the following thirty-five variables were used for this research:

- Eight variables were clinical or demographic. They included age at study entry, sex (male/female), number of days on study, study arm, disease site (breast/gynecological/colorectal), previous chemotherapy (yes/no), information if the disease was metastatic or non-metastatic, and the number of comorbidities.

- Twenty-four variables were from PROMs questionnaires, completed by participants on paper at the time of study entry (baseline):
 - Fifteen of these variables were from EORTC QLQ-C30 (apetite, dyspnoea, pain, fatigue, nausea and vomiting, constipation, diarrhea, finance, global health, cognitive function, sleep, emotional well-being, role well-being, physical well-being, and social well-being). These were the calculated scores for the 15 aspects of health-related quality of life from all 30 items. [45]
 - Five were from EQ-5D (social well-being, mobility, self-care, usual activities, pain, and anxiety and depression) [202]
 - Four were from FACT-G (physical, social, emotional, and functional well-being) [46]

The aims of chapter include exploring the effects of all PROMs on the model performance in general, and when specific questionnaires are separately added to clinical data. Therefore, 6 different variable sets were created:

1. **Only clinical variables:** to be used as a baseline for the comparison of performance of models with other variable sets;
2. **All available variables:** to explore the general effect of adding all available PROMs data to clinical data on the model performance;
3. **Only PROMs variables;** to investigate if PROMs with no clinical information can be used for prediction;
4. **Clinical variables and QLQ-C30 variables:** to investigate if QLQ-C30 questionnaire specifically adds predictive value compared to other PROMs questionnaires;
5. **Clinical variables and EQ-5D variables:** to investigate if EQ-5D questionnaire specifically adds predictive value compared to other PROMs questionnaires;
6. **Clinical variables and FACT-G variables:** to investigate if FACT-G questionnaire specifically adds predictive value compared to other PROMs questionnaires.

Target Variables

Three target variables in the dataset were the number of unplanned hospital admissions, the number of triage episodes (patients contacting the emergency helpline or attending the oncology admission unit with symptoms) and changes to chemotherapy (any alteration of medication or a dose) during 18 weeks (entire duration) of the clinical trial. These variables were transformed to binary variables, where class 0 indicates no events of admissions, triage episodes or chemotherapy during the trial, and class 1 indicates at least one admission, triage episode or change to chemotherapy. This variable transformation was performed to enable binary classification and is commonly done on outcome variables in studies using ML for health outcome predictions [106, 109].

4.2.3 Data Pre-processing

Missing Data Handling

Figure 4.2 illustrates the number of available (non-missing) values in each variable. As described in Chapter 3, section 3.3, all patients were expected to complete all three PROMs questionnaires: QLQ-C30, EQ-5D and FACT-G. However, there were 91 participants who took part in the initial pilot phase of the eRAPID clinical trial and completed only two subscales of QLQ-C30. The complete version of QLQ-C30 questionnaire was added in the full trial and collected from all remaining patients. This explains the large difference in the number of the values missing in QLQ-C30 variables, compared to all other variables. Any further cases of missing values were infrequent (Figure 4.2) and likely resulted from participants omitting questions while completing questionnaires, which is a common issue in PROMs data [215].

Due to the different reasons for missing data, they were handled in two ways:

1. The records from the 91 participants who did not complete the full version of QLQ-C30 questionnaires were removed from the affected variable sets. Using Complete Case (CC) analysis in this instance is justified, as the participants whose records were removed are from a pilot trial, and therefore randomly selected. Under the Missing Completely at Random (MCAR) assumption that participants whose data are deleted do not differ from the remaining participants, the CC method is unlikely to bias the results [216].
2. Minor cases of missing data were imputed using KNN ($k=5$). This method of data imputation was chosen due to its frequent use in previous studies using machine learning for outcome prediction in health [6, 102, 106, 107]. It is also considered as an accurate form of data imputation technique [182].

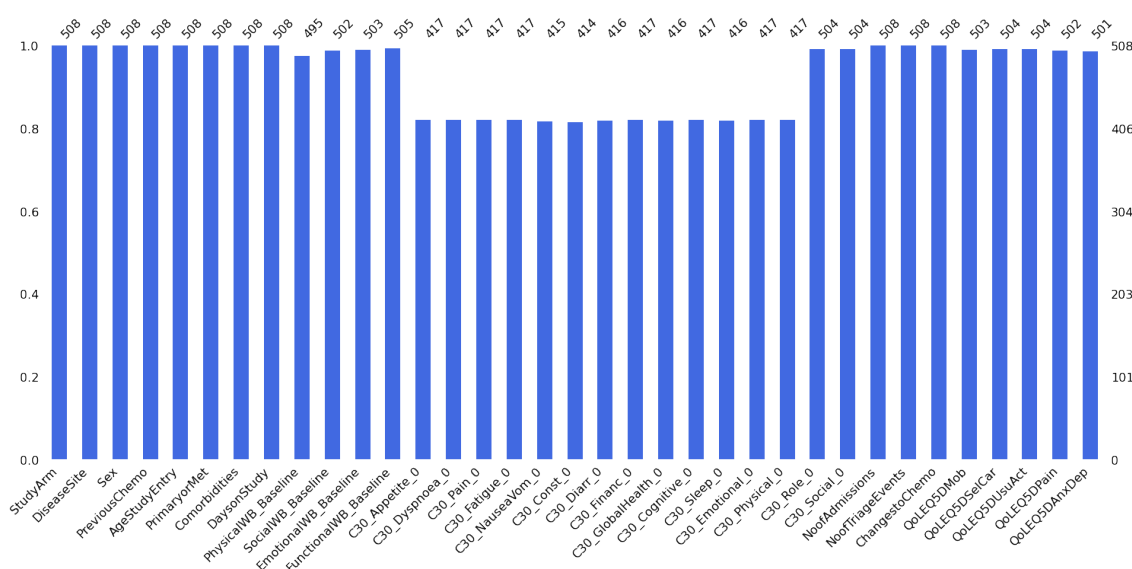


Figure 4.2: Number of non-missing data in each variable.

Variable Preparation

Many ML methods require standardisation of continuous variables in order to improve computational performance [217]. Therefore, the data were scaled to unit variance. Furthermore, it is a common practice to remove correlation between input variables. It can aid the preparation of ML models through preventing algorithms from receiving repeated information [218]. Pairs of variables which had the Pearson correlation coefficient of over 0.6 were identified. The same threshold was used in previous studies exploring variable selection methods for classification tasks [219]. The correlated variables were removed from each of the 6 variable sets, leaving one out of the pairs or sets of correlated variables (Table 4.1).

Class Imbalance

There was a class imbalance in all three prediction targets. More patients experienced at least one triage episode ($n=294$) than no triage episodes ($n=214$); more patients had their chemotherapy changed at least once ($n=333$) than not changed ($n=175$); and more patients were not admitted to the hospital during the trial at all ($n=338$) than were admitted at least once ($n=170$). Table 4.2 presents the differences between classes for admissions, triage and changes to chemotherapy.

Class imbalance is one of the most pressing challenges in medical datasets, which could lead to biased ML models [12]. Studies predicting patient outcomes from PROMs have mitigated this issue in various ways. For example, models can be trained on original data and evaluated using multiple performance metrics [102]. However, class imbalance can be addressed using a pre-processing technique called oversampling. This method creates new synthetic participants in the minority class to match the number of participants in the majority class. In previous studies this oversampling has been done either on the entire dataset [6, 97] or on train set only, leaving the test set imbalanced [19, 93]. Undersampling is another method which can handle class imbalance, but it could reduce valuable information from the data in small datasets. Therefore, oversampling is a better solution in this research.

Since there is no consistency in the methodologies of existing research predicting patient outcomes from PROMs, all three methods were applied. In order to evaluate if handling class imbalance method can affect the results, three separate datasets were created from each of the six variable sets and considering each of the three target variables:

1. **Original dataset:** No pre-processing method was done to address class imbalance. 80% of the data were used to train the models and 20% were used as a test set. This method might affect the model performance, as it might be biased towards the majority class [183].
2. **Balanced dataset:** Random sampling with replacement (oversampling) was performed on the entire dataset. The train/test split with 8:2 ratio was done after the data were balanced. This method assumes that the model will be tested on data that is pre-processed. Therefore, the real-world relevance might be affected.

3. **Partially balanced dataset:** Dataset was first split to train/test sets with 8:2 ratio. The split was stratified, ensuring the same proportion of classes in both sets, as in the entire dataset [106]. Random sampling with replacement was then performed on the 80% of the data (train set) and the test set remained imbalanced. According to literature, this method provides the opportunity for the model to learn well from the data, without pre-processing the test set [101].

4.2.4 Model Development

Six ML models were selected as classifiers in this chapter, based on their inclusions in previous research predicting patient outcomes from PROMs [6, 19]. Hyperparameter tuning was performed on the train sets through grid search with five-fold cross-validation. Model with the parameters leading to the best performance was used for evaluation on the test sets, which is a common practice in similar research [19]. The evaluation metric used to select the parameters in the development stage was Area Under the Receiver Operating Characteristic Curve (AUC), which is the most common metric in similar research (Chapter 2). The following models were applied:

1. **Logistic Regression (LR):** A binary classifier estimating regression equation for a target variable [6]. Hyperparameters considered for LR were regularisation strength and solver.
2. **Decision Tree (DT):** A model which uses nodes and branches to perform splitting through choosing between input variables based on a selected parametric criterion [6]. Hyperparameters considered were criterion, maximum depth, the minimum number samples required to split and minimum number of samples required to be at a leaf node.
3. **Adaptive Boosting (AB):** An ensemble classification model combining different classifiers to complement their weak points[6]. Hyperparameters considered were number of estimators, learning rate and boosting algorithm.
4. **Random Forest (RF):** Another ensemble classifier using decision trees on various generated train sets and uses averaging to improve the predictive performance [6]. Hyperparameters considered for RF were criterion, maximum depth, the minimum number samples required to split and the number of estimators.
5. **Support Vector Machines (SVM):** A model which uses complex variable space transformation to discriminate between classes applying a hyperplane [19]. Hyperparameters considered were kernel type, kernel coefficient and regularisation parameter.
6. **Multilayer Perceptron (MLP):** A model that mimics human brain through interconnected neurons. It includes an input layer, hidden layers, and an output layer [19]. Hyperparameters considered were activation function for the hidden layer, solver for weight optimization, the number of neurons in hidden layers, and the strength of the L2 regularization term.

Table 4.1: Variables in each variable sets. The highlighted bold variables were selected to be included in the models, and the light variables were removed based on the high correlation coefficient.

1. Clinical variables	2. All variables	3. PROMs	4. Clinical + QLQ-C30	5. Clinical + EQ-5D	6. Clinical + FACT-G
'DiseaseSite'	'DiseaseSite'	'PhysicalWB_Baseline'	'DiseaseSite'	'DiseaseSite'	'DiseaseSite'
'Sex'	'Sex'	'SocialWB_Baseline'	'Sex'	'Sex'	'Sex'
'PreviousChemo'	'PreviousChemo'	'EmotionalWB_Baseline'	'PreviousChemo'	'PreviousChemo'	'PreviousChemo'
'AgeStudyEntry'	'AgeStudyEntry'	'FunctionalWB_Baseline'	'AgeStudyEntry'	'AgeStudyEntry'	'AgeStudyEntry'
'PrimaryorMet'	'PrimaryorMet'	'QLQ-C30_Appetite_0'	'PrimaryorMet'	'PrimaryorMet'	'PrimaryorMet'
'Comorbidities'	'Comorbidities'	'QLQ-C30_Dyspnoea_0'	'Comorbidities'	'Comorbidities'	'Comorbidities'
'DaysonStudy'	'DaysonStudy'	'QLQ-C30_Pain_0'	'DaysonStudy'	'DaysonStudy'	'DaysonStudy'
'StudyArm'	'StudyArm'	'QLQ-C30_Fatigue_0'	'StudyArm'	'StudyArm'	'StudyArm'
	'PhysicalWB_Baseline'	'QLQ-C30_Nausea Vom_0'	'QLQ-C30_Appetite_0'	'QoLEQ5DMMob'	'PhysicalWB_Baseline'
	'SocialWB_Baseline'	'QLQ-C30_Const_0'	'QLQ-C30_Dyspnoea_0'	'QoLEQ5DSeICar'	'SocialWB_Baseline'
	'EmotionalWB_Baseline'	'QLQ-C30_Diarr_0'	'QLQ-C30_Pain_0'	'QoLEQ5DUsuAct'	'EmotionalWB_Baseline'
	'FunctionalWB_Baseline'	'QLQ-C30_Financ_0'	'QLQ-C30_Fatigue_0'	'QoLEQ5DDPain'	'FunctionalWB_Baseline'
	'QLQ-C30_Appetite_0'	'QLQ-C30_GlobalHealth_0'	'QLQ-C30_Nausea Vom_0'	'QoLEQ5DAnxDep'	
	'QLQ-C30_Dyspnoea_0'	'QLQ-C30_Cognitive_0'	'QLQ-C30_Const_0'		
	'QLQ-C30_Pain_0'	'QLQ-C30_Sleep_0'	'QLQ-C30_Diarr_0'		
	'QLQ-C30_Fatigue_0'	'QLQ-C30_Emotional_0'	'QLQ-C30_Financ_0'		
	'QLQ-C30_Nausea Vom_0'	'QLQ-C30_Physical_0'	'QLQ-C30_GlobalHealth_0'		
	'QLQ-C30_Const_0'	'QLQ-C30_Role_0'	'QLQ-C30_Cognitive_0'		
	'QLQ-C30_Diarr_0'	'QLQ-C30_Social_0'	'QLQ-C30_Sleep_0'		
	'QLQ-C30_Financ_0'	'QoLEQ5DMMob'	'QLQ-C30_Emotional_0'		
	'QLQ-C30_GlobalHealth_0'	'QoLEQ5DSeICar'	'QLQ-C30_Physical_0'		
	'QLQ-C30_Cognitive_0'	'QoLEQ5DUsuAct'	'QLQ-C30_Role_0'		
	'QLQ-C30_Sleep_0'	'QoLEQ5DDPain'	'QLQ-C30_Social_0'		
	'QLQ-C30_Emotional_0'	'QoLEQ5DAnxDep'			
	'QLQ-C30_Physical_0'				
	'QLQ-C30_Role_0'				
	'QLQ-C30_Social_0'				
	'QoLEQ5DMMob'				
	'QoLEQ5DSeICar'				
	'QoLEQ5DUsuAct'				
	'QoLEQ5DDPain'				
	'QoLEQ5DAnxDep'				

Table 4.2: Differences in variables between classes in all target variables and the p-value (* $P \leq .01$, ** $P \leq .001$, *** $P \leq .05$) generated by Mann-Whitney U test for all variables but AgeStudyEntry (t-test used). Mean (M), counts of values for categorical variables, and mean for continuous variables are reported.

	Not admitted (n=338)		Admitted (n=170)		Significance test		Not triaged (n=214)		Triaged (n=294)		Significance test		No chemotherapy change (n=175)		Chemotherapy change (n=333)		Significance test	
	M	Mean/Value	M	Mean/Value			M	Mean/Value	M	Mean/Value			M	Mean/Value	M	Mean/Value		
StudyArm	1	1=70, 2=108	1	1=76, 2=94	0.475		2	1=104, 2=110	1	1=152, 2=142	0.245		2	1=82, 2=93	1	1=174, 2=159	0.124	
DiseaseSite	2	1=137, 2=58, 3=123	2	1=76, 2=48, 3=46	0.220		2	1=85, 2=49, 3=80	1	1=148, 2=57, 3=89	0.011*		2	1=122, 2=21, 3=32	1	1=111, 2=85, 3=137	<0.001	
Sex	2	F=263, M=75	2	F=143, M=27	0.047*		2	F=162, M=52	2	F=214, M=50	0.022*		2	F=150, M=25	2	F=256, M=77	0.009**	
PreviousChemo	0	No=296, Yes=72	0	No=131, Yes=39	0.337		0	No=155, Yes=59	0	No=242, Yes=52	0.004**		0	No=145, Yes=30	0	No=252, Yes=81	0.032*	
AgeStudyEntry	56	55.429	57.5	56.994	0.158		57	56.972	55	55.21	0.097		54	54.126	58	56.913	0.011*	
PrimarytMet	1	1=228, 2=110	1	1=89, 2=81	<0.001***		1	1=32, 2=82	1	1=185, 2=109	0.388		1	1=128, 2=47	1	1=189, 2=144	<0.001***	
Comorbidities	0	0=192, 1=103, 2=36, 3=7	1	0=75, 1=54, 2=30, 3=11	<0.001***		0	0=119, 1=62, 2=25, 3=8	0	0=148, 1=95, 2=41, 3=10	0.136		0	0=108, 1=52, 2=14, 3=1	1	0=159, 1=105, 2=52, 3=17	<0.001***	
DaysonStudy	126	119.891	126	119.888	0.118		126	119.575	126	120.119	0.489		126	120.211	126	119.721	0.085	
PhysicalWB_Baseline	25	23.725	24	22.424	0.003**		25	23.711	24	22.980	0.071		25	23.522	24.5	23.015	0.039**	
SocialWB_Baseline	24	23.693	24	23.552	0.126		24	23.451	24	23.788	0.305		25	24.325	24	23.295	<0.001***	
FunctionalWB_Baseline	18	16.769	17	16.429	0.131		18	16.872	18	16.497	0.173		18	16.692	18	16.653	0.346	
FunctionalWB_Baseline	19.917	19.016	19	18.263	0.123		20	19.088	19	18.527	0.154		20.5	19.572	19	18.346	0.010*	
QLQ-C30 Appetite.0	0	16.606	0	27.381	0.115		0	17.356	0	22.361	0.043*		0	13.043	0	23.775	0.041*	
QLQ-C30 Dyspnoea.0	0	8.424	0	20.000	0.003**		0	10.358	0	13.750	0.090		0	9.179	0	13.859	0.291	
QLQ-C30 Pain.0	16.667	20.818	16.667	25.833	0.124		16.667	19.868	16.667	24.444	0.057		16.667	21.981	16.667	22.760	0.377	
QLQ-C30 Fatigue.0	22.222	26.835	33.333	36.984	0.002**		22.222	26.962	33.333	32.662	0.012*		22.222	25.644	33.333	32.517	0.089	
QLQ-C30 Nausea Vom.0	0	3.579	0	9.643	0.028*		0	4.571	0	6.736	0.012*		0	3.406	0	7.014	0.205	
QLQ-C30 Const.0	0	12.774	0	15.952	0.160		0	15.048	0	12.971	0.334		0	10.706	0	15.403	0.337	
QLQ-C30 Diarr.0	0	9.179	0	10.000	0.443		0	7.910	0	10.600	0.043*		0	8.213	0	10.072	0.379	
QLQ-C30 Financ.0	0	15.644	0	13.810	0.246		0	15.443	0	14.722	0.358		0	16.667	0	14.217	0.155	
QLQ-C30 GlobalHealth.0	75	73.853	75	67.679	0.037*		83.333	74.479	75	69.792	0.122		83.333	75.302	75	70.024	0.014*	
QLQ-C30 Cognitive.0	83.333	83.514	83.333	80.595	0.250		83.333	85.217	83.333	80.556	0.159		83.333	83.454	83.333	82.079	0.221	
QLQ-C30 Sleep.0	33.333	37.681	33.333	37.619	0.388		33.333	35.028	33.333	39.699	0.033*		33.333	36.983	33.333	37.993	0.294	
QLQ-C30 Emotional.0	73	73.345	83.333	73.000	0.261		83.333	76.177	75	72.257	0.227		75	74.155	75	73.865	0.274	
QLQ-C30 Physical.0	93.333	84.073	80	76.976	0.012*		86.667	82.420	86.667	81.153	0.493		93.333	86.993	86.667	79.068	<0.001***	
QLQ-C30 Role.0	83.333	73.980	83.333	70.513	0.103		83.333	75.708	83.333	70.719	0.046*		83.333	76.686	66.667	70.796	0.032*	
QLQ-C30 Social.0	83.333	74.527	83.333	72.189	0.224		83.333	75.236	83.333	72.660	0.234		83.333	77.168	83.333	71.954	0.029**	
QLQ-C30 Mob	1	1.369	1	1.547	0.032*		1	1.458	1	1.409	0.314		1	1.276	1	1.511	<0.001***	
QLQ-E03DScCar	1	1.141	1	1.241	0.056		1	1.292	1	1.155	0.244		1	1.103	1	1.212	0.045*	
QLQ-E03DScAct	1	1.731	2	1.847	0.107		1	1.746	1	1.757	0.289		1	1.693	2	1.858	0.002**	
QLQ-E03DScPain	2	1.810	2	1.947	0.101		2	1.829	2	1.876	0.309		2	1.776	2	1.889	0.046*	
QLQ-E03DScDep	2	1.807	2	1.894	0.183		2	1.762	2	1.890	0.060		2	1.805	2	1.853	0.194	

Python *sklearn* library was used to apply all ML models. Each model was fitted to predict each of the 3 target variables, based on each of the 6 variable sets for original, balanced and partially balanced data. In total, outcomes of 324 ML models (for 6 algorithms, 3 outcomes, 6 variable sets, in 3 types of data) were evaluated.

4.2.5 Model Evaluation

The metrics used to evaluate model performances were accuracy, precision, recall, F1-score, and AUC. Each value was obtained through evaluating the models with hyperparameters established in grid search on a test set. Accuracy is a proportion of correctly identified cases. Precision is the number of instances correctly classified to a particular class divided by the total number of instances classified to this class. Recall is the number of instances correctly classified to a particular class divided by the total number of instances belonging to that class. F1 score aggregates precision and recall through the harmonic mean of these metrics [220]. AUC is most frequently reported as a main performance measure and is calculated as the area under a curve plotted from sensitivity against false positive rate (1-specificity). The wide range of performance metrics used allows in depth understanding of the models and compare results with other similar studies. AUC is considered the main performance metric to allow comparisons between models.

Model calibrations were evaluated with calibration plots of RF in balanced datasets and LR in remaining datasets because of the best overall performance of these models in these scenarios. Analysis of Variance (ANOVA) with Tukey's Honest Significant Difference (Tukey HSD) tests were performed to compare model performances. Variable importance analyses were also performed on these models. LR variables were analysed through the absolute values of regression coefficients. This method is only meaningful for standardized data with no multicollinearity [221], which was accounted for by standardisation of variables and removing correlated variables. RF variables were explored through "variable importances" python command in *sklearn.RandomForestClassifier*. [222].

4.2.6 Stakeholder Engagement

The methodology described in this section was designed through both the review of the relevant literature (Chapter 2) and engagement with clinicians and health scientists. The methods and preliminary results of this chapter were presented to PCOR Group in the University of Leeds, Faculty of Medicine and Health, Leeds Institute of Medical Research. Initially, only original and balanced datasets were taken into consideration in this study. However, the researchers from the PCOR group reported concerns regarding the reliability of the model performance in balanced datasets, as the models were tested on balanced (non-real) data. This issue is also known in literature to affect any possibility of the clinical adoption of ML models. Therefore, the partially balanced dataset was added to the analyses, enabling testing models on real data. Furthermore, clinicians were interested in which information models were considering when predicting outcomes, so variable importance analyses were added to this chapter.

4.3 Results

The results (performance metric values for each model), and hyperparameters selected through grid search for each model, are presented in appendix Figure A5. The AUC values are plotted for hospital admission (Figure 4.3), triage (Figure 4.5), and changes to chemotherapy (Figure 4.6).

4.3.1 Hospital Admissions

Overall Predictive Value of PROMs. This paragraph explores if variable sets including PROMs led to better performance than clinical and demographic variables only, in all 3 scenarios for addressing class imbalance.

1. For all models in original dataset, clinical variables had worse accuracy, precision, recall, F1 score and AUC than any of the feature sets including PROMs, indicating that PROMs add predictive value to the ML models.
2. As in original dataset, all performance metrics were higher for variable sets including PROMs than for only clinical variables.
3. In the partially balanced dataset, F1 score was higher for clinical variables in SVM (0.188) and NN (0.493) than for other variable sets. Nevertheless, these values were not the highest overall, achieved by other models. For SVM, recall was also the highest value for the clinical variables (0.176). No AUC was the highest for clinical variables, which also suggest better overall performance of models including PROMs than only clinical variables.

Individual PROMs Questionnaires. This paragraph explores the predictive value of individual PROMs questionnaires:

1. In the original dataset, clinical + QLQ-C30 variables achieved the best AUC in all models except NN, for which AUC was highest for PROMs only. This suggests that for most models, QLQ-C30 questionnaire improves model performance more than other questionnaires.
2. In the balanced dataset, clinical + QLQ-C30 variables obtained the highest AUC in all models, apart from LR and SVM, where AUCs were highest for all variables. This also suggests a good predictive value of QLQ-C30.
3. In the partially balanced dataset, the highest AUC was obtained by clinical + QLQ-C30 variables in LR, DT, RF (the same value as PROMs only), and NN. In AB, the highest AUC was achieved by all variables, and in SVM by clinical + FACT-G variables. Overall, QLQ-C30 variables aided ML performance the most.

Model Performance. This paragraph compares the performance of ML models in all scenarios:

1. In the original dataset, LR performed best in all variable sets, except for clinical + EQ-5D, where DT was superior (Figure 4.3). The highest AUC (0.659) was obtained by LR with clinical + QLQ-C30 variables.

2. In the balanced dataset, RF was the best performing algorithm (highest AUC = 0.905) for all variable sets, apart from all variables and clinical + EQ-5D variables, where SVM performed slightly better.
3. In the partially balanced dataset, the best AUC (0.616) was achieved by LR using clinical + QLQ-C30 variables and AB using all variables.

Balancing the entire dataset improved model performance based on all evaluation metrics. Using partially balanced data resulted in similar AUCs and precision to original data, but improved F1 score and recall. Model calibration for predicting admissions was poor in original data and improves slightly in balanced and partially balanced scenarios (Figure 4.4).

4.3.2 Triage Events

Overall Predictive Value of PROMs. Assessing if PROMs add predictive value to ML models in all scenarios:

1. The highest values obtained by only clinical variables in original data were F1 score (0.771) and recall (1.00) for NN. Overall model performance assessed by AUC was always the highest for variable sets including PROMs.
2. In balanced data only precision (0.585) of AB model was the highest for clinical variables only, but all other metrics were the highest for variables including PROMs.
3. No AUC was the highest for clinical variables only in any of the models predicting triage in partially balanced scenario.

Individual PROMs questionnaires. Assessing the predictive value of individual PROMs questionnaires:

1. For predicting triage, in original dataset the feature sets that most frequently achieved the highest AUCs were: only PROMs variables for LR, DT and RF; all variables for SVM and NN; and clinical + QLQ-C30 variables for AB.
2. In balanced dataset, all variables obtained the highest AUC for DT, AB, RF and SVM. Clinical + QLQ-C30 variables had the highest performance for LG, and clinical + EQ-5D variables for NN.
3. In partially balanced data, clinical + EQ-5D variables had the best performance the most frequently (for AB, RF, SVM and NN), and clinical + FACT-G variables feature set was selected twice (for LR and DT).

Due to the lack of consistency between the performance of models built on different variable sets, no single questionnaire can be recommended.

Model Performance. Comparison of the performance of ML models in all scenarios.

1. LR is the best performing model in predicting triage events from all feature sets, apart from PROMs only variables in original data, where the best AUC of 0.596 was achieved by DT.

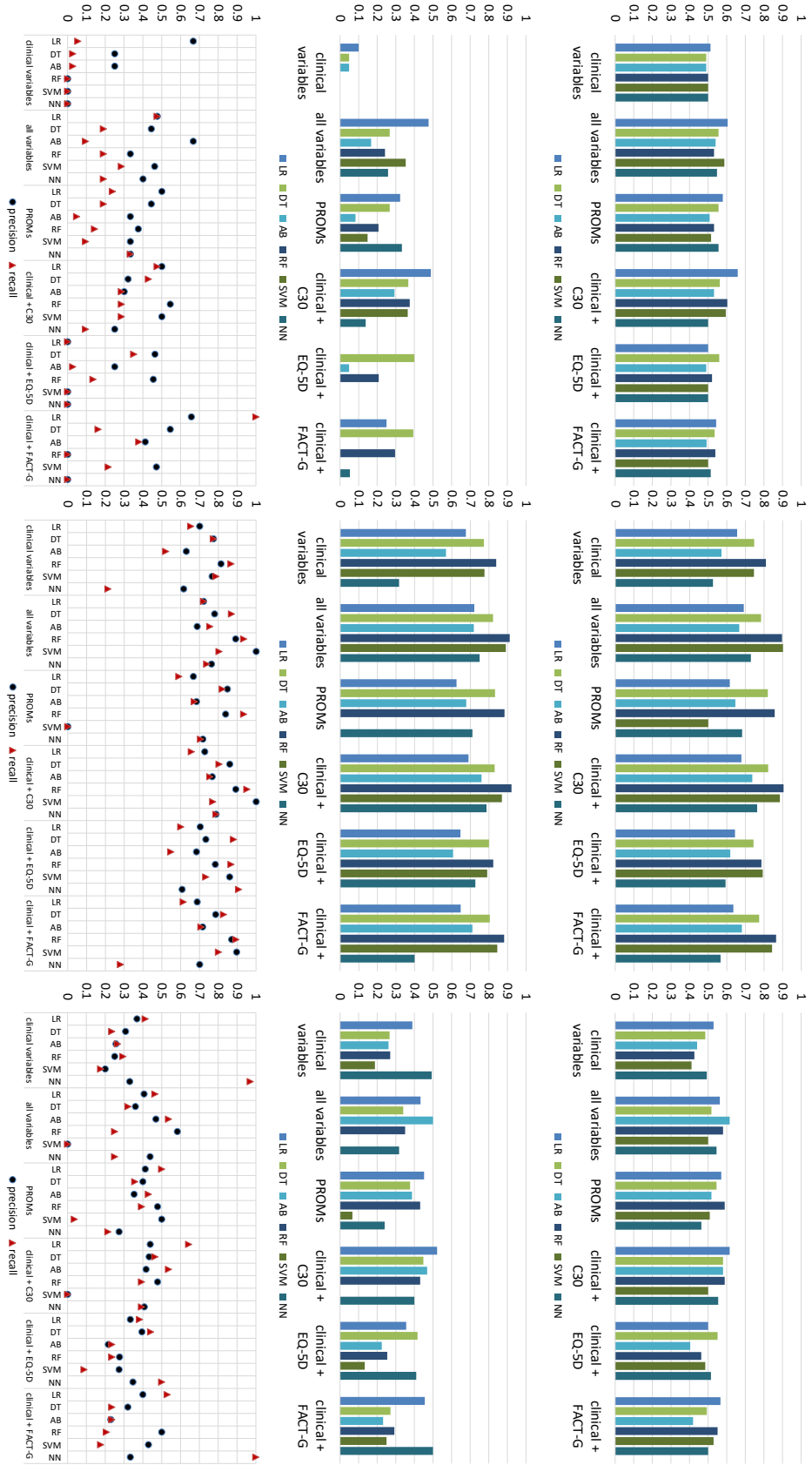


Figure 4.3: AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting admissions.

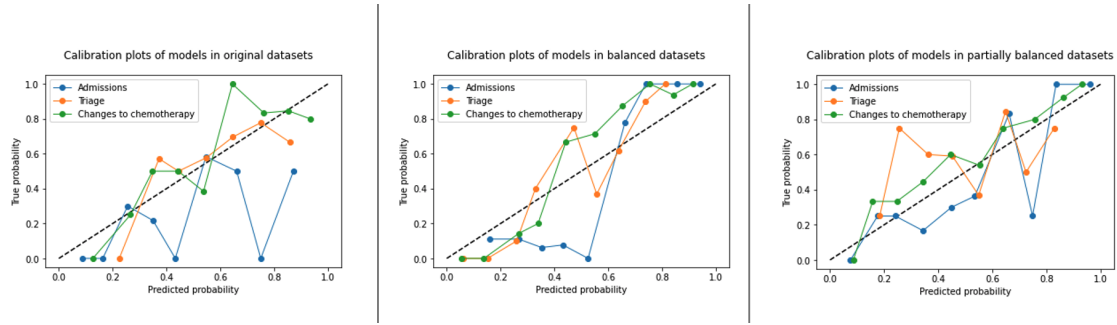


Figure 4.4: Calibration plots of LR models in original dataset (left), calibration of RF models in balanced datasets (middle), and calibration of LR models in partially balanced datasets (right). LR, logistic regression, RF, random forest.

2. In balanced dataset the best performance is achieved by SVM or RF with the highest AUC of 0.764 for SVM predicting triage from all variables.
3. Like for admissions, there is no obvious best algorithm for predicting triage in partially balanced data, but the best AUC of 0.624 was obtained by LR in clinical + FACT-G feature set.

A slight increase in the AUCs for the balanced data set is noticeable in Figure 4.5. AUCs and F1 scores in the partially balanced data set were similar to the original data set. For some models in balanced data, the F1 scores were lower than in other data sets. Model calibration remained poor across different data types (Figure 4.4).

4.3.3 Changes to Chemotherapy

Overall Predictive Value of PROMs. Assessing if PROMs add predictive value to ML models in all scenarios:

1. In the original dataset, AUC was the highest for clinical variables only in DT (0.623) and RF (0.623). However, these values were not the highest overall. Other models achieved higher AUC for models including PROMs.
2. In the balanced dataset, recall was the only measure highest for clinical variables in AB (0.754) and NN (1.00). Overall performance was better for variable sets including PROMs.
3. In the partially balanced dataset, clinical variables obtained the highest precision (0.872) and AUC (0.682) in LR and SVM, respectively. Overall, highest AUC had models including PROMs.

Individual PROMs questionnaires. Assessing predictive value of individual PROMs questionnaires:

1. When predicting chemotherapy changes, in original data clinical + QLQ-C30 variables had the highest performance in LR and SVM, and clinical + FACT-G variables for AB and NN.

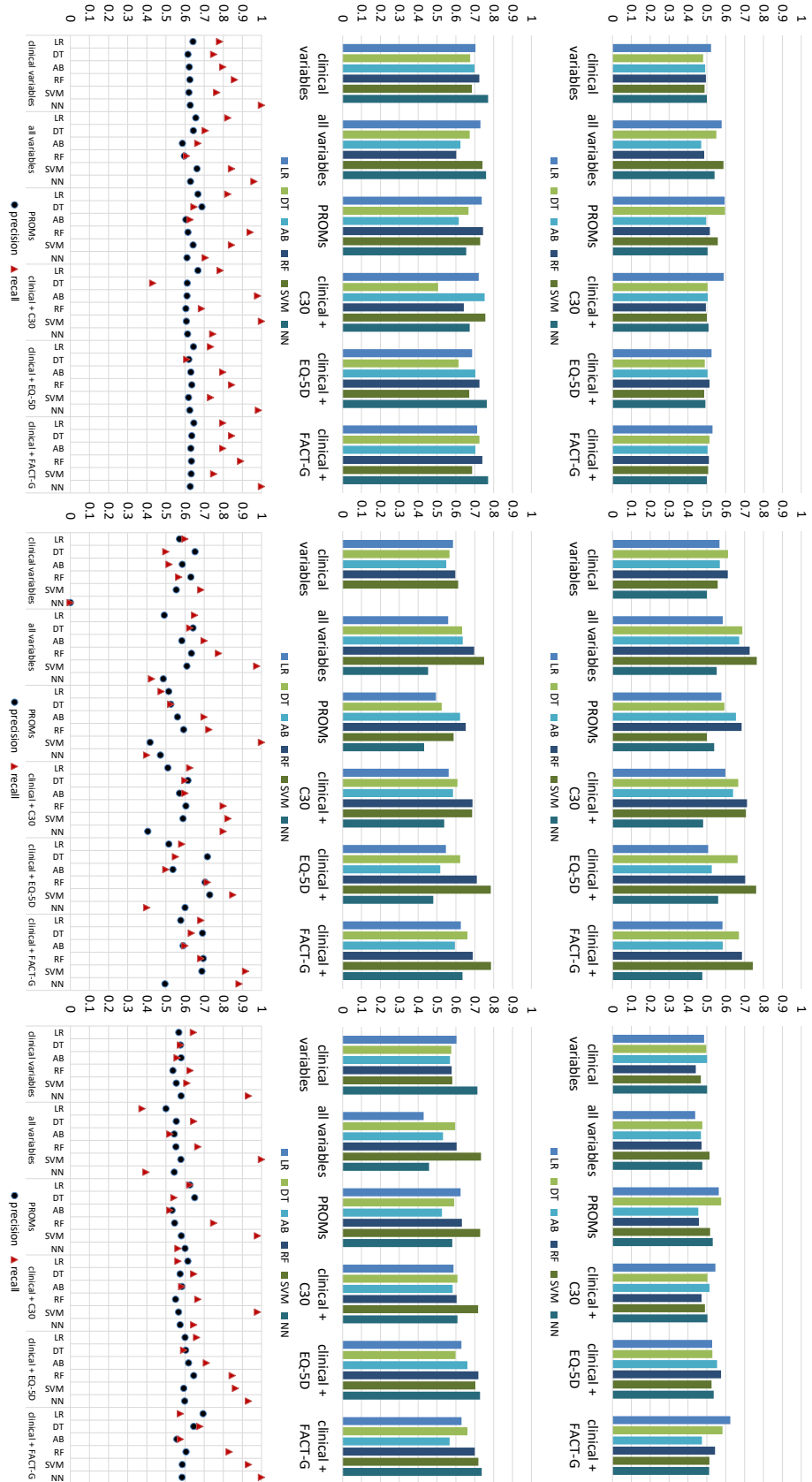


Figure 4.5: AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting triage.

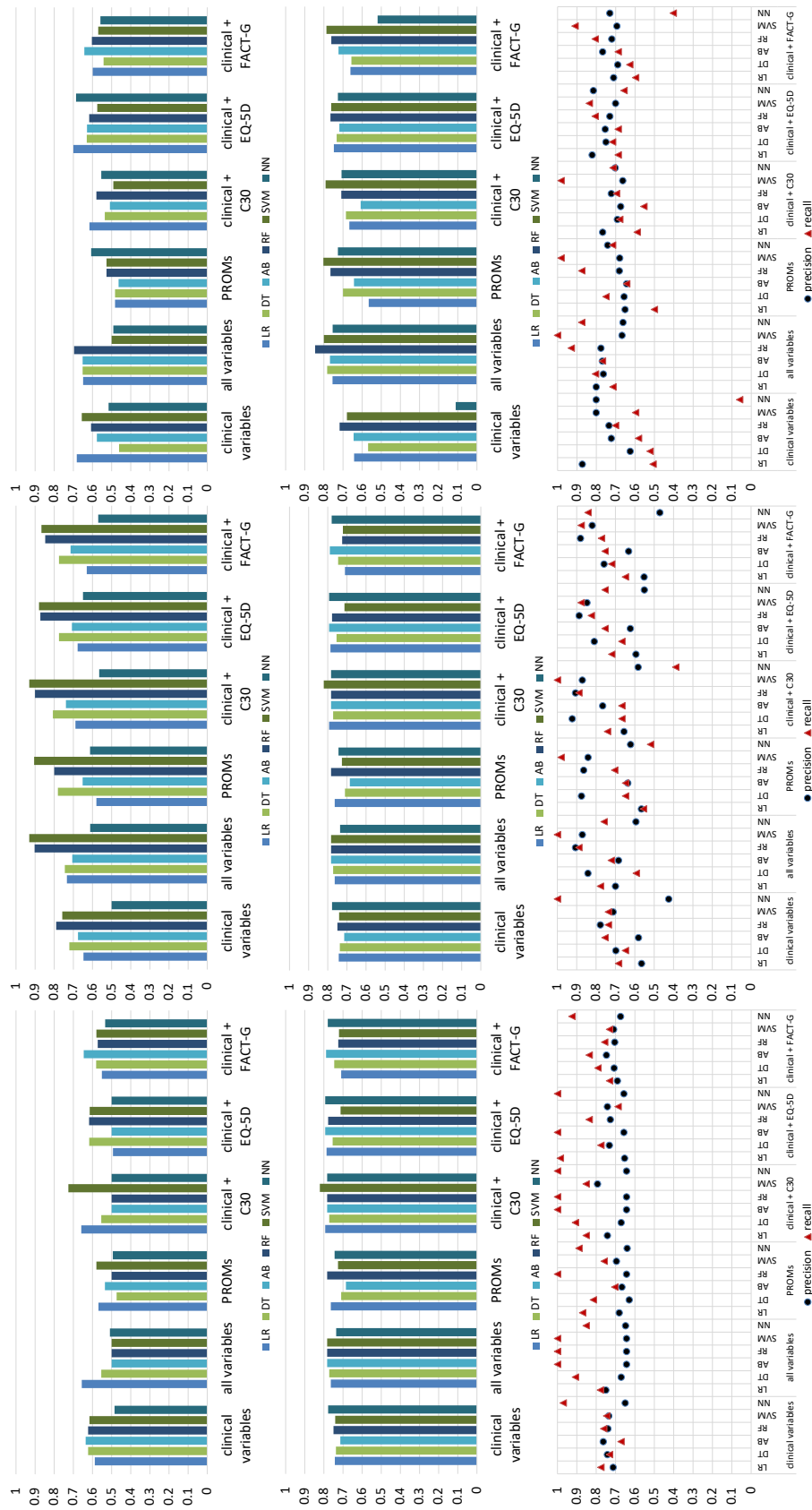


Figure 4.6: AUC values (top row), F1 scores (middle row) and precision and recall distributions (bottom row) for original (left column), balanced (middle column) and partially balanced (right column) datasets when predicting changes to chemotherapy.

2. In the balanced dataset, clinical + QLQ-C30 variables obtained the highest AUC in DT and AB, clinical + EQ-5D variables for NN, and all variables for LR and RF. In SVM, all variables and clinical + QLQ-C30 variables achieved the same AUC (0.931).
3. In the partially balanced dataset, all variables obtained the highest AUC for DT, AB, and RF; only clinical variables for LR and SVM; and clinical + EQ-5D variables for NN.

These results also do not indicate conclusive predictive value of any of the questionnaire.

Model Performance.

1. In the original dataset, all models achieved very similar performance.
2. In the balanced dataset, the best algorithms were RF and SVM. SVM with all and clinical + QLQ-C30 variables had the best overall performance (AUC = 0.931).
3. In the partially balanced dataset, similarly to the original dataset, no one model had an outstanding performance.

There was no noticeable difference between original and partially balanced datasets (Figure 4.6). Model calibration was very good in the partially balanced dataset and slightly worse in other data types (Figure 4.4).

4.3.4 Significance Testing

Overall, models predicting changes to chemotherapy performed significantly better than models predicting triage ($P < .01$) and admissions ($P < .001$). ANOVA also suggests that performances from various models differed significantly ($F=4.351$, $P<0.001$). However, Tukey HSD did not reveal any significant differences between individual models, apart from DT performing better than NN ($P < 0.05$), and SVM performing better than NN ($P < .01$). ANOVA also suggests that the results differ significantly based on balancing methods, with balanced dataset leading to significantly higher performance than original ($P < 0.05$) and partially balanced ($P < 0.001$) datasets.

4.3.5 Variable Importance

In predicting admissions, LR prioritized clinical variables, while RF focused on PROMs and age at study entry (Table 4.3). The feature importance analysis was added following the consultation with PCOR group. In predicting triage, LR mainly focused on clinical variables with individual FACT-G and EQ-5D variables. RF prioritised FACT-G and some clinical variables (Table 4.3). In predicting changes to chemotherapy, LR prioritized clinical variables with individual FACT-G and EQ-5D variables. RF mainly considered FACT-G and some clinical variables (Table 4.3). These variable importance analysis for each target variables were very similar across all scenarios with the models developed in all types of data balancing techniques.

Table 4.3: Variable importance for LR (coefficient) and RF (value) models predicting all 3 target variables. The lower the rank of the variable, the higher the importance for a particular model.

			Original data				Balanced data				Partially balanced data			
			Logistic regression		Random Forest		Logistic regression		Random Forest		Logistic regression		Random Forest	
			coef	rank	value	rank	coef	rank	value	rank	coef	rank	value	rank
Admissions	Clinical/demographic	AgeStudyEntry	0.005	19	0.052	5	-0.041	13	0.093	3	0.019	12	0.108	1
		StudyArm	0.060	11	0.006	23	0.100	12	0.013	22	0.085	9	0.021	20
		DiseaseSite	-0.348	4	0.019	17	-0.412	2	0.033	13	-0.404	3	0.032	15
		PreviousChemo	-0.402	2	0.003	24	-0.290	4	0.012	24	-0.241	4	0.012	23
		PrimaryorMet	0.866	1	0.037	13	1.047	1	0.044	9	0.732	1	0.033	12
		Comorbidities	0.366	3	0.034	14	0.335	3	0.039	10	0.484	2	0.055	6
		DaysonStudy	0.007	17	0.037	12	0.004	19	0.019	20	0.006	18	0.013	22
	QLQ-C30	QLQ-C30_Appetite_0	0.008	16	0.106	3	0.011	15	0.050	5	0.008	16	0.036	10
		QLQ-C30_Dyspnoea_0	0.013	14	0.128	2	0.013	14	0.048	7	0.017	14	0.042	8
		QLQ-C30_NauseaVom_0	0.013	15	0.052	6	-0.003	24	0.017	21	0.006	19	0.017	21
		QLQ-C30_Const_0	-0.003	20	0.009	21	-0.005	17	0.026	17	-0.002	23	0.033	14
		QLQ-C30_Diarr_0	-0.006	18	0.013	20	-0.003	20	0.023	19	-0.004	21	0.024	19
		QLQ-C30_Financ_0	-0.001	22	0.040	10	-0.007	16	0.030	15	-0.007	17	0.032	16
		QLQ-C30_Cognitive_0	-0.001	23	0.030	15	-0.003	22	0.045	8	0.003	22	0.043	7
		QLQ-C30_Sleep_0	-0.003	21	0.014	19	-0.003	21	0.035	12	-0.005	20	0.035	11
		QLQ-C30_Social_0	0.000	24	0.039	11	0.005	18	0.048	6	-0.013	15	0.061	5
	EQ-5D	QoLEQ5DMob	-0.061	10	0.017	18	0.225	8	0.029	16	-0.002	24	0.030	17
		QoLEQ5DSeIcar	0.197	6	0.055	4	0.112	11	0.013	23	-0.217	5	0.009	24
		QoLEQ5DUsuAct	0.165	8	0.043	9	-0.147	9	0.024	18	0.124	7	0.033	13
		QoLEQ5DPain	0.030	13	0.045	8	0.128	23	0.031	14	-0.132	6	0.029	18
		QoLEQ5DAnxDep	-0.039	12	0.009	22	0.239	7	0.037	11	-0.117	8	0.038	9
	Fact-G	PhysicalWB_Baseline	-0.188	7	0.132	1	-0.260	5	0.104	1	-0.017	13	0.094	2
		FunctionalWB_Baseline	0.163	9	0.051	7	0.126	10	0.102	2	-0.079	10	0.090	3
		SocialWB_Baseline	0.248	5	0.028	16	0.259	6	0.086	4	0.053	11	0.080	4
Triage	Clinical/demographic	AgeStudyEntry	-0.127	10	0.097	2	-0.186	8	0.119	1	-0.265	3	0.121	1
		StudyArm	-0.162	9	0.006	23	-0.202	7	0.014	22	0.243	4	0.026	17
		DiseaseSite	-0.166	8	0.013	22	-0.255	5	0.027	16	-0.196	7	0.030	15
		PreviousChemo	-0.721	1	0.019	19	-0.430	3	0.013	24	-0.643	1	0.022	20
		PrimaryorMet	0.487	2	0.004	24	0.644	1	0.018	20	0.624	2	0.017	23
		Comorbidities	0.338	5	0.044	10	0.397	4	0.036	11	0.108	9	0.036	10
		DaysonStudy	0.012	15	0.034	12	0.001	24	0.014	23	0.010	19	0.020	21
	QLQ-C30	QLQ-C30_Appetite_0	0.006	18	0.077	3	0.002	23	0.032	12	0.000	24	0.031	14
		QLQ-C30_Dyspnoea_0	0.003	23	0.028	16	0.004	21	0.026	17	0.015	16	0.019	22
		QLQ-C30_NauseaVom_0	-0.005	20	0.031	13	0.009	16	0.019	19	0.002	22	0.022	19
		QLQ-C30_Const_0	-0.006	17	0.016	20	-0.010	15	0.032	13	-0.017	15	0.034	13
		QLQ-C30_Diarr_0	0.004	21	0.053	8	0.008	17	0.029	14	0.026	12	0.036	9
		QLQ-C30_Financ_0	-0.004	22	0.019	18	-0.006	19	0.026	18	-0.010	20	0.029	16
		QLQ-C30_Cognitive_0	-0.005	19	0.062	6	0.002	22	0.043	7	-0.009	21	0.044	7
		QLQ-C30_Sleep_0	0.006	16	0.050	9	0.005	20	0.049	5	0.012	18	0.051	5
		QLQ-C30_Social_0	-0.002	24	0.029	15	-0.008	18	0.049	6	-0.001	23	0.044	6
	EQ-5D	QoLEQ5DMob	0.024	13	0.013	21	-0.183	10	0.028	15	-0.115	8	0.026	18
		QoLEQ5DSeIcar	-0.444	3	0.042	11	-0.207	6	0.014	21	-0.222	6	0.008	24
		QoLEQ5DUsuAct	0.283	6	0.056	7	0.088	14	0.037	9	0.021	13	0.034	12
		QoLEQ5DPain	0.053	12	0.030	14	0.440	2	0.040	8	0.019	14	0.035	11
		QoLEQ5DAnxDep	0.069	11	0.019	17	0.100	13	0.037	10	0.013	17	0.037	8
	Fact-G	PhysicalWB_Baseline	-0.258	7	0.122	1	-0.131	12	0.096	4	-0.234	5	0.091	3
		FunctionalWB_Baseline	0.364	4	0.074	4	0.186	9	0.103	2	0.074	10	0.096	2
		SocialWB_Baseline	0.013	14	0.063	5	0.171	11	0.099	3	0.036	11	0.090	4
Changes to chemotherapy	Clinical/demographic	AgeStudyEntry	-0.295	6	0.113	2	-0.138	6	0.112	2	-0.179	9	0.127	1
		StudyArm	-0.167	9	0.008	24	-0.137	7	0.016	22	-0.084	13	0.021	19
		DiseaseSite	1.007	1	0.128	1	-0.197	4	0.098	3	0.755	1	0.071	5
		PreviousChemo	0.034	12	0.012	23	-0.159	5	0.016	21	-0.226	6	0.015	22
		PrimaryorMet	-0.723	2	0.022	17	0.294	3	0.018	19	-0.299	3	0.018	20
		Comorbidities	0.572	3	0.064	6	0.316	2	0.031	12	0.396	2	0.038	10
		DaysonStudy	-0.002	21	0.024	16	0.000	24	0.014	23	0.003	23	0.009	24
	QLQ-C30	QLQ-C30_Appetite_0	0.010	17	0.035	8	0.003	22	0.039	8	0.011	17	0.029	14
		QLQ-C30_Dyspnoea_0	0.000	24	0.020	19	0.004	21	0.013	24	0.005	22	0.016	21
		QLQ-C30_NauseaVom_0	0.009	18	0.019	21	0.008	16	0.021	18	0.037	15	0.021	19
		QLQ-C30_Const_0	0.004	20	0.027	14	-0.010	15	0.026	14	-0.008	20	0.029	15
		QLQ-C30_Diarr_0	-0.010	16	0.022	18	0.008	17	0.023	16	-0.011	18	0.028	16
		QLQ-C30_Financ_0	-0.010	15	0.029	12	-0.006	19	0.032	11	-0.009	19	0.034	11
		QLQ-C30_Cognitive_0	0.000	23	0.024	15	0.001	23	0.038	9	0.014	16	0.044	7
		QLQ-C30_Sleep_0	0.001	22	0.044	7	0.005	20	0.040	7	0.006	21	0.041	9
		QLQ-C30_Social_0	0.006	19	0.034	9	-0.007	18	0.046	6	-0.001	24	0.047	6
	EQ-5D	QoLEQ5DMob	0.241	7	0.033	10	-0.135	9	0.027	13	0.137	10	0.027	17
		QoLEQ5DSeIcar	0.446	4	0.019	20	-0.129	10	0.017	20	0.282	4	0.013	23
		QoLEQ5DUsuAct	0.052	11	0.032	11	0.043	14	0.034	10	0.218	7	0.041	8
		QoLEQ5DPain	-0.211	8	0.029	13	0.334	1	0.024	15	-0.128	11	0.031	13
		QoLEQ5DAnxDep	-0.136	10	0.016	22	0.063	13	0.023	17	-0.056	14	0.031	12
	Fact-G	PhysicalWB_Baseline	-0.011	14	0.066	5	-0.125	11	0.064	5	0.274	5	0.079	4
		FunctionalWB_Baseline	-0.026	13	0.112	3	0.108	12	0.143	1	-0.091	12	0.109	2
		SocialWB_Baseline	-0.358	5	0.071	4	0.136	8	0.085	4	-0.215	8	0.082	3

4.4 Discussion

4.4.1 Overall Results

Predictive Value of PROMs

Overall, the AUC values suggest that PROMs improved performance of ML models predicting hospital utilisation in this study in all scenarios. This improvement in performance of models including PROMs is compatible with research suggesting that HRQOL can enhance cancer survival predictions [6]. It is also consistent with the study using EHR and PROMs to predict hospital utilisation in patients during chemotherapy, where PROMs added some, but limited predictive value [223]. In general, for admissions and changes to chemotherapy in all datasets, it is apparent that F1 score also improved with PROMs additions. Looking at the distribution of precision and recall, it was mainly the precision values that influenced that difference, as they were usually the lowest for clinical variables. This difference is especially noticeable in the original dataset when predicting admissions, and original and partially balanced dataset when predicting changes to chemotherapy (figures 4.3 and 4.6). This suggests that PROMs might have enabled identifying patients who were admitted to hospital or had changes in chemotherapy more accurately than those who did not experience these events. There have also been cases where models with only PROMs variables outperformed models which included clinical variables. Although there is evidence suggesting that using PROMs without objectively measured data in ML models can lead to accurate predictions of patient outcomes [73], there should be more studies exploring datasets with exclusively PROMs. However, the variable importance suggests that models tend to focus on clinical variables. This might suggest that including these variables might be beneficial.

Individual PROMs Questionnaires

Although adding all PROMs variables to the models generally improved model performance, it was very beneficial to look into 3 different questionnaires separately, as variable sets with only one questionnaire added to clinical variables often performed better than all variables. The QLQ-C30 questionnaire seemed to add the most predictive value to all of the methods. QLQ-C30 was the only questionnaire with variables consistently significantly different between classes (4.1), which might have improved the model performances. This effect was especially noticeable for admissions prediction. Since QLQ-C30 questionnaire is most frequently used to measure quality of life in cancer patients [224], these results are promising, as the availability of the questionnaire may aid its utilisation in ML models for clinical practice.

Model Performance

Overall, LR was the best performing model in original datasets predicting admissions and triage. In changes to chemotherapy, the performance was dependent on variable sets. In partially balanced data there was also no outstanding model. Nevertheless, in balanced data for all target variables, RF and SVM performed much better than other models. The AUC values achieved by these models when predicting admissions and changes to chemotherapy in balanced data were over 0.9, suggesting a very good performance of this model, compatible with ensemble methods used by Sim et al.,

[6] and higher than models used by Peterson et al., [223] and Sidey-Gibbons et al., [188]. LR being the simplest model and outperforming other ML algorithms on imbalanced data was also observed in previous studies [97, 106].

Effects of Balancing Data

Balancing the entire dataset before splitting into test and train set resulted in the best performance, which was very high for admissions and changes to chemotherapy predictions (AUC higher than 0.9), and considerably good for triage predictions (the highest AUC=0.764). Unfortunately, the promising results from the balanced dataset might affect generalisability of the models, as upsampling often causes overfitting [225]. Therefore, any potential application in clinical practice might not be possible with such method, which was also the concern of clinicians consulted about the study design. Nevertheless, training models on imbalanced data can lead to incorrect predictions, biased towards one of the classes [226], which was apparent through low recall values in admissions predictions, being the most imbalanced target variable.

Another method explored in this study was training models on balanced data and testing on a 20% hold-out set. There were no noticeable differences for any of the metrics, compared with original data in triage target variable. This might be due to lower level of class imbalance, compared to other outcome variables (Table 4.2). In admissions and changes to chemotherapy there was a slight improvement in performance for F1 score and AUC values. The partially balanced dataset mitigated the bias of model evaluation on pre-processed data, and the lack of generalisability, while still proving the best (balanced) environment for model training. Upsampled train set can ensure the robustness of models, and testing on original hold-out set could provide more accurate perspective for real clinical data [101].

Variable Importance

In all target variables, LR models tended to focus more on clinical variables than on PROMs. RF models usually favoured FACT-G variables with some relevant clinical or QLQ-C30 variables (mainly sleep, cognitive and social scales). These patterns were similar for all target variables. LR was often the best-performing model (according to AUC) in original data, which could explain the poor performance of predicting triage and admissions, as the clinical variables for these targets did not have significant differences between classes (Table 4.2), yet the model was considering these variables as the most important (Table 4.3). For changes to chemotherapy, there were a lot of significantly different clinical variables, which might have aided the good performance of LR. RF performing well on balanced data would suggest the significance of FACT-G questionnaire, contrary to the results of its variables in model performances. The inconsistency between the variable importance analysis and the predictive value of individual PROMs may be the effect of different sizes of the datasets. FACT-G, in comparisons with other sets has the fewest number of variables, so the questionnaire did not contribute as much as QLQ-C30. Nevertheless, its individual variables may have played an important role in predictions where included in the analyses.

4.4.2 Strengths and Limitations

Strengths

This work provides important insights into PROMs included in ML models to predict patient outcomes and highlights the issues in conducting and reporting research.

1. **PROMs questionnaires used:** Three different validated PROMs questionnaires were used and their impact was analysed through different variable sets and individual variable importance. This allowed the exploration of which aspects of patients' reported data could be useful when making treatment choice and aid any potential application of ML methods in clinical practice based on PROMs availability.
2. **Techniques for data pre-processing compared:** This study addresses the inconsistency in imbalance data handling pre-processing methods in existing studies [6, 19, 93, 97, 102] and informs about significant differences in results generated from 3 techniques. This difference provide the extremely important insight into how pre-processing methods can bias results. It recommends model evaluation on unprocessed, real-world data to improve clinical relevance of research which struggle with class imbalance.
3. **Performance metrics used:** This study also used the variety of performance metrics, which allowed between-studies comparison of PROMs predictive value [73] and in-depth understanding of the models. Although reporting more than one evaluation metric is common, the analyses of precision and recall allowed exploration of weak sides of each model.
4. **Clinician involvement:** Another strength of this research is the engagement with the PCOR group (which comprised 3-4 oncologists, cancer nurses and health scientists with extensive expertise in using PROMs in oncology) during the methods development. According to the scoping review described in Chapter 2, there is no clinical involvement reported in most studies predicting patient outcomes from PROMs. Involving health professionals, even in early stages of exploratory research, could support the clinical adoption of various AI methods [227].

Limitations

Nevertheless, this work also has several limitations.

1. **Representativeness of the dataset:** The study uses data collected in a clinical trial which might not be representative of the general population due to exclusion and inclusion criteria [228]. Therefore, external validation of this research would strengthen the conclusions. However, considering that this is a case study, aiming to explore the predictive value of PROMs and bias caused by pre-processing methods, it provides the necessary comparison of different scenarios.
2. **Clinical trial design:** Half of the participants (intervention group) was using the intervention, which might have affected the outcome. Furthermore, random forest including all variables predicting admissions, triage and chemotherapy changes on balanced and imbalanced data was run and tested on a) only

control group data; b) only intervention group data. The largest difference between the values of AUC was 0.05. Based on these results and the fact that the clinical trial also did not find the intervention group to be significantly different from the control group for most of the outcomes, the potential bias of the intervention group can be rejected.

3. **Sample size:** The sample size used for the analysis in this chapter was relatively small. Small sample size is associated with higher accuracy in studies using ML classifiers which might suggest model overfitting [229]. This work presented the importance of inclusion PROMs in ML models, which should encourage further collection of patient-reported data, and hopefully result in larger and more insightful datasets. Therefore, in further chapters this work will also focus on longitudinal PROMs data, as changes in reporting over time might provide more meaningful conclusions.
4. **Stakeholder engagement:** Even through clinician involvement indirectly represented the views of patients, there were no consultations with any patient representatives during the design of this chapter. For this stage of the thesis, organising a session with a PCOR group was feasible. However, due to logistical issues and the time needed to invite stakeholders and organise sessions, no patients representatives were consulted. Wider stakeholder sessions will take place as a part of this work, involving patients, as well as clinicians, ensuring the feasibility of the research and seeking design feedback.

4.4.3 Conclusions

This chapter predicted hospital utilisation from PROMs completed at the baseline of chemotherapy and compared there methods for handling class imbalance. It started exploring the answers to the following research questions:

1. **What is the added value of using PROMs in AI models predicting patient outcomes?** This question was answered through comparing models applied on clinical variables with the variable sets that included and excluded PROMs. It further explored the predictive value of individual PROMs questionnaires and the variable importance of particular variables. The results supported the evidence that PROMs improve performance of models predicting patient outcomes, which in this case was hospital utilisation. Therefore, the results further motivate PROMs collection, which can benefit both shared decision making process in clinical practice, and ML research predicting patient outcomes. This study also provides the insight into the predictive value of particular PROMs questionnaires, and encourages the use of widely available PROMs, such as QLQ-C30 questionnaire.
2. **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** Model performances were compared in three different scenarios: original data, balanced data (prior to train/test split), and partially balanced data (balanced train set, without affecting test set). Based on large discrepancy across results from different pre-processing methods, this research alerts the scientific community to justify choices of methods for handling class imbalance and understand how these can affect the results. It is

recommended to balance the train set only and to test models on original data to prevent bias.

3. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** The consultation with PCOR group and the collaboration with a clinical oncologist throughout the design of this study provided meaningful insights into clinical relevance of this chapter. The design was strengthened by ensuring test set is not affected by the balancing methods and adding feature importance analysis. Nevertheless, this research question will further be explored in this thesis, including the involvement of patient representatives in the study design.

This chapter focused on baseline PROMs for the predictions of hospital utilisation outcomes, providing insights into PROMs' predictive value, data pre-processing methods, and stakeholder engagement. The further chapters in this thesis will build on the findings reported in this chapter. The next chapter will similarly use baseline PROMs, but the predicted outcomes will be selected by patient representatives and focus on HRQoL measures. The work described in further chapters will also use the recommendation to protect the test set from pre-processing methods to mimic real-world scenarios when evaluating models. Furthermore, the model selections in the further chapters will build on the results presented here, by removing models which were not successful in processing the data for prediction of hospital utilisation.

Chapter 5

Using Baseline PROMs To Predict HRQoL Changes

5.1 Introduction

The previous chapter provided an insight into using PROMs collected at one time point to predict hospital utilisation during chemotherapy treatment. The study design was consulted with clinicians to understand the clinical relevance of this research and the concerns regarding methodology. However, the views of patients were not considered in the design of the research conducted so far. Therefore, this chapter aims to bridge that gap through addressing the following research questions:

1. **What is the added value of using PROMs in AI models predicting patient outcomes?** Since the previous chapter focused on PROMs as input data in AI models predicting patient outcomes, this chapter utilises PROMs as both input and target variables. Health-Related Quality of Life (HRQoL) measures are rarely used in AI predictive modelling research, so this chapter explores the prediction of improvement and deterioration of PROMs values at different stages of chemotherapy.
2. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** The patient representatives involved in this research helped understand what aspects of chemotherapy were unexpectedly difficult. They mentioned that physical symptoms and hospital utilisation are important outcomes, but patients are expecting to experience them throughout the chemotherapy. However, the quality of life changes throughout the treatment were often missed from the conversations leading to treatment decisions, even though these outcomes affect patients' well-being. Therefore, this study focuses on predicting HRQoL during chemotherapy.

Patients' views can be considered in two different ways in research predicting healthcare outcomes. They can be represented through structured PROMs used as input data for predictive modelling, as well as through consultations with patients throughout the study design. Cancer prediction and diagnosis models rarely include patient-reported data [230], missing the information on patients' perspective on their health. Including PROMs in models predicting patient outcomes can increase the relevance of the input data and provide more insights

from the model to learn from. As the previous chapter suggests, PROMs also have a potential to improve the model performance. However, there is little evidence of patient or clinician engagement during the study design in the literature, as explored in the scoping review (Chapter 2). Since the development of AI models is not based on patients' perspective or clinical knowledge, it might not actually serve its purpose or address needs of diverse group of people in an equal and fair way. Involving patients and public input is essential not only for the feasibility of novel models, but also for increasing public trust in AI research [37]

Cancer diagnosis can be devastating for an individual and lead to emotional distress, anxiety and depression [231]. Patients are usually faced with the immediate need to make a decision regarding their treatment, which is a daunting and complicated process. Often the choice of treatment requires the compromise between the quality of life and the length of life. However, the preferences and values of patients vary depending on the individual differences [232]. For instance, quality of life is usually an extremely important factor for older patients, whilst younger patients are more likely to undergo more aggressive treatment to increase survival years [233]. Therefore, it is crucial to take individual perspective on the personal values and characteristics in the decision-making process. Unfortunately, there are limited available sources informing prognosis, choices, alternatives and consequences of cancer treatment. This often leads to patients deciding to undergo chemotherapy, even when it has very little benefit to their health, can cause major side effects, and as a result drastically decrease the quality of life [234].

There are existing studies exploring DL models for predictions of oncology outcomes, but the main focus has been on toxicity symptoms [88, 208, 235]. Studies foreseeing HRQoL during chemotherapy tend to use statistical modelling, rather than explore ML predictive opportunities [236]. Furthermore, the outcomes of ML models are usually limited to specific time during the chemotherapy, without investigating the changes from early to late stages of the treatment [88]. Patient and clinician involvement is also missing from the design of existing studies, as reported in Chapter 2. Therefore, this study uses PROMs data to predict the quality of life changes from baseline to 6, 12, and 18 weeks of chemotherapy. This outcome has been chosen as meaningful by patients' consulted during the design stage of this project. Therefore, the research described in this chapter contributes to understanding how patients can be involved in research using PROMs and AI for the prediction of patient outcomes. The methodology follows the recommendations from the research reported in the previous chapter.

5.2 Methods

5.2.1 Methodology for Predicting Health-Related Quality of Life Changes from Baseline PROMs

Four machine learning models: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP) were used to predict patient-reported HRQoL changes at 6, 12 and 18 weeks of starting chemotherapy treatment. The outcomes included physical well-being, social

functioning, role functioning, usual activities, and mobility. These outcomes were consulted with patients’ representatives who mentioned that changes in HRQoL are difficult to prepare for and can impact the well-being more than hospitalisation. Therefore, the outcomes represent the information which is relevant to patients. The process of stakeholder engagement is further described in Section 5.2.6. The overall study design is presented in Figure 5.1.

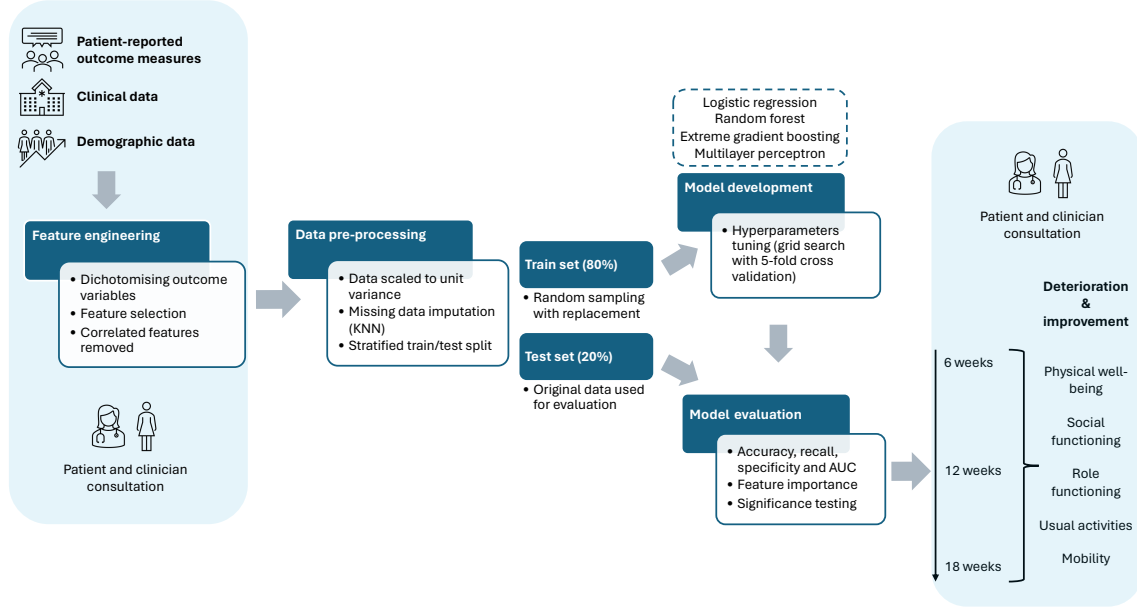


Figure 5.1: Flow diagram illustrating the overall methodology of the study.

5.2.2 Dataset

Input Variables

The dataset contains data from 508 patients of the eRAPID clinical trial [70], described in Chapter 3. The clinical and demographic variables remained the same as in the previous study. However, the variables indicating all individual comorbidities presented in Table 3.2 in Chapter 3 were added. Therefore, 42 variables are demographic or clinical. The remaining 48 variables consist of PROMs, completed by participants at 4 time points (baseline at the start of the trial, and at 6-, 12-, and 18-week follow-up). For each time point, 6 variables were from EQ-5D [202]); and 4 variables from FACT-G [46]), as in the previous study.

The patients in the pilot of the eRAPID clinical trial only completed two variables from QLQ-C30: social and role functioning scale [45], as reported in previous chapter (Figure 4.2). Only these two variables from the QLQ-C30 questionnaire were included in the analysis for this chapter. This is different to the previous chapter, where all the variables remained in the dataset, but the patients who did not complete all QLQ-C30 questions were excluded. In this instance, the number of patients’ data points change as the outcomes is measured at different time-points. Due to loss to follow-up, there are limited number of patients whose data can be analysed, so removing patients’ data further would result in very small

datasets. Furthermore, the QLQ-C30 variables correlate to variables from the other questionnaires, so removing them does not compromise the information in data, which prevents missing data affecting the results. All variables used in this chapter are presented in Table 5.3.

Target Variables

All outcomes predicted in this study were consulted with patients in order to ensure that the research is clinically relevant and addresses the real challenge patients undergoing chemotherapy are facing. Contrary to our predictions, consulted patients emphasised that they are well informed about the expected impact of chemotherapy on their physical symptom. However, they would like to understand what is the expected impact on their daily life, represented through the HRQoL measures. The target variables were improvement and deterioration in: physical well-being (physical symptoms), social functioning (ability to engage in the society), role functioning (ability to perform life roles), usual activities (ability to perform daily life activities), and mobility (ability to move) at 6, 12 and 18 weeks of chemotherapy. To establish the change of the HRQoL measures (improvement or deterioration), the baseline values were subtracted from the values at the predicted time point. These values indicating the magnitude and direction of change were transformed into binary variables to indicate the improvement (1) or no improvement, including no change or deterioration (0), and deterioration (1) or no deterioration, including no change or improvement (0). The way of the transformation depended on the type of questionnaire which the outcome was taken from. For physical well-being, social functioning and role functioning the threshold of MCID was used, as the clinical relevance of these changes differed based on the scale. The MCID values were taken from previous studies, where they were established for each variable [237, 238]. Usual activities and mobility collected with EQ-5D were 5-level, 1-item scale (from no problems to extreme problems) and have no MCID recommendations. Therefore, the deterioration and improvement of these variables were considered any change of ≤ 1 and ≤ -1 respectively. This is informative, as for 5-item questionnaires each unit suggests change. The magnitude of change for each outcome, and the questionnaires used for their collections are presented in Table 5.1.

Table 5.1: Predicted variables, questionnaires completed, change calculations, participants number (N), including deterioration (D) and improvement (I) cases.

HRQOL	Questionnaire	Deterioration	Improvement	Time point (N, D, I)
Physical well-being	Fact-G	change ≤ -2	change ≥ 2	6 weeks (N=439, D=252, I=68)
				12 weeks (N=400, D=241, I=51)
				18 weeks (N=382, D=219, I=58)
Social functioning	QLQ-C30	change ≤ -7	change ≥ 8	6 weeks (N=440, D=187, I=101)
				12 weeks (N=407, D=209, I=73)
				18 weeks (N=388, D=179, I=90)
Role functioning	QLQ-C30	change ≤ -6	change $\geq 11^*$	6 weeks (N=438, D=190, I=100)
				12 weeks (N=407, D=216, I=90)
				18 weeks (N=385, D=171, I=78)
Usual activities	EQ-5D	change ≥ 1	change ≥ -1	6 weeks (N=439, D=148, I=71)
				12 weeks (N=408, D=181, I=50)
				18 weeks (N=384, D=183, I=51)
Mobility	EQ-5D	change ≥ 1	change ≥ -1	6 weeks (N=438, D=64, I=48)
				12 weeks (N=408, D=98, I=40)
				18 weeks (N=382, D=122, I=31)

*MCID for role functioning improvement was the average from all cancer types due to availability [238].

Table 5.2: Removed variables based on correlation analysis.

Physical well-being	6 weeks	'SOC0', 'ROL0', 'AgeStudyEntry', 'Sex', 'EWB_overall0'
	12 weeks	'FWB_overall0', 'SOC0', 'ROL0', 'AgeStudyEntry', 'Sex', 'EWB_overall0', 'CMRheuCTD'
	18 weeks	'AgeStudyEntry', 'EWB_overall0', 'Sex', 'SOC0'
Social functioning	6 weeks	'SOC0', 'EWB_overall0', 'AgeStudyEntry', 'ROL0', 'Sex'
	12 weeks	'SOC0', 'AgeStudyEntry', 'EWB_overall0', 'CMRheuCTD', 'CMCarHyperten', 'FWB_overall0', 'ROL0', 'Sex'
	18 weeks	'SOC0', 'EWB_overall0', 'AgeStudyEntry', 'Sex'
Role functioning	6 weeks	'SOC0', 'ROL0', 'Sex', 'AgeStudyEntry', 'FWB_overall0', 'EWB_overall0'
	12 weeks	'SOC0', 'ROL0', 'Sex', 'CMRheuCTD', 'AgeStudyEntry', 'FWB_overall0', 'EWB_overall0', 'CMCarHyperten'
	18 weeks	'SOC0', 'Sex', 'AgeStudyEntry', 'EWB_overall0', 'CMCarHyperten'
Usual activities	6 weeks	'ROL0', 'EWB_overall0', 'FWB_overall0', 'Sex', 'AgeStudyEntry', 'SOC0'
	12 weeks	'ROL0', 'EWB_overall0', 'CMCarHyperten', 'FWB_overall0', 'CMRheuCTD', 'Sex', 'AgeStudyEntry', 'SOC0'
	18 weeks	'EWB_overall0', 'FWB_overall0', 'Sex', 'AgeStudyEntry', 'SOC0'
Mobility	6 weeks	'AgeStudyEntry', 'PWB_overall0', 'SOC0', 'ROL0', 'Sex', 'FWB_overall0', 'EWB_overall0'
	12 weeks	'AgeStudyEntry', 'ROL0', 'SOC0', 'Sex', 'CMCarHyperten', 'FWB_overall0', 'CMRheuCTD', 'EWB_overall0'
	18 weeks	'AgeStudyEntry', 'SOC0', 'ROL0', 'Sex', 'EWB_overall0'

5.2.3 Data Pre-Processing

As in the previous chapter, variables correlated with Pearson coefficient higher than 0.6 were removed (leaving one) to ensure ML models process information efficiently. The list of removed variables in each variable set is presented in the Table 5.2. Continuous variables were scaled to unit variance to improve computational performance of models [217]. As data were collected in a clinical trial, inevitably there was a loss to follow up, which resulted in different number of participants who completed PROMs at each time point [70]. Only rows (participants) with complete target variable at the predicted time point were included in the analysis. The exact numbers of participants are presented in Table 5.1.

Other minor cases of missing data were imputed using KNN imputer (k=5) [239], as in the previous chapter. The number of missing data for each variable in each outcome is presented in Table 5.3. Original class distribution is provided in Table 5.1. As recommended in previous chapter, only train set was affected by random sampling with replacement to ensure that models are not biased towards one class [226]. Test set was left imbalanced which enabled models to be evaluated on unprocessed data to increase the clinical relevance and a potential to be applied to clinical practice. The data was split to train (80%) and test (20%) with stratification, in order to retain the proportion of classes in each set [106].

5.2.4 Model Development

The models used in this chapter slightly differed from the models used in the previous chapter. LR, RF, and MLP were used here as well due to their frequency of use in studies applying ML to PROMs data to enable between-studies comparison, as established in the scoping review. SVM and DT were not used in this chapter, as their overall performance in the previous chapter was not better than other models' performances. Furthermore, they are less frequently used in other research, as established in the scoping review, reported in Chapter 2. Furthermore, AB was substituted for another boosting method: Extreme Gradient Boosting (XGB), as it often outperforms other ML models [240] in similar studies and has a potential to achieve better performance than AB. Therefore, LR, RF, XGB and MLP were applied using Python sklearn library to predict the improvement and deterioration of each HRQoL outcome at 6, 12, and 18 weeks from baseline. Hyperparameters were tuned on train sets through grid search with five-fold cross-validation. The

Table 5.3: Names of all variables used in the research presented in this chapter and the number of missing data in each variable for each outcome at 18 weeks.

Physical well-being		Social functioning		Role functioning		Usual activities		Mobility	
StudyArm	0	StudyArm	0	StudyArm	0	StudyArm	0	StudyArm	0
DiseaseSite	0	DiseaseSite	0	DiseaseSite	0	DiseaseSite	0	DiseaseSite	0
Sex	0	Sex	0	Sex	0	Sex	0	Sex	0
PreviousChemo	0	PreviousChemo	0	PreviousChemo	0	PreviousChemo	0	PreviousChemo	0
AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0
PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0
BCBMI	1	BCBMI	1	BCBMI	1	BCBMI	1	BCBMI	2
Comorbidities	0	Comorbidities	0	Comorbidities	0	Comorbidities	0	Comorbidities	0
CMCarMI	0	CMCarMI	0	CMCarMI	0	CMCarMI	0	CMCarMI	0
CMCarAngina	0	CMCarAngina	0	CMCarAngina	0	CMCarAngina	0	CMCarAngina	0
CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0
CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0
CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0
CMCarVenous	0	CMCarVenous	0	CMCarVenous	0	CMCarVenous	0	CMCarVenous	0
CMResCOPD	0	CMResCOPD	0	CMResCOPD	0	CMResCOPD	0	CMResCOPD	0
CMResEmphys	0	CMResEmphys	0	CMResEmphys	0	CMResEmphys	0	CMResEmphys	0
CMResAsthma	0	CMResAsthma	0	CMResAsthma	0	CMResAsthma	0	CMResAsthma	0
CMResChronBron	0	CMResChronBron	0	CMResChronBron	0	CMResChronBron	0	CMResChronBron	0
CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0
CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0
CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0
CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0
CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0
CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0
CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0
CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0
CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0
CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0
CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0
CMNeuMS	0	CMNeuMS	0	CMNeuMS	0	CMNeuMS	0	CMNeuMS	0
CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0
CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0
CMRheuArth	0	CMRheuArth	0	CMRheuArth	0	CMRheuArth	0	CMRheuArth	0
CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0
CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0
CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0
CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0
CMPrevMal	0	CMPrevMal	0	CMPrevMal	0	CMPrevMal	0	CMPrevMal	0
CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0
CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0
DCMarital	4	DCMarital	4	DCMarital	4	DCMarital	4	DCMarital	4
DCEmployment	12	DCEmployment	12	DCEmployment	12	DCEmployment	12	DCEmployment	12
ed_lev	12	ed_lev	12	ed_lev	12	ed_lev	12	ed_lev	12
EQ5DMob0	4	EQ5DMob0	4	EQ5DMob0	4	EQ5DMob0	3	EQ5DMob0	2
EQ5DSelCar0	3	EQ5DSelCar0	3	EQ5DSelCar0	3	EQ5DSelCar0	2	EQ5DMob18	0
EQ5DUusuAct0	3	EQ5DUusuAct0	3	EQ5DUusuAct0	3	EQ5DUusuAct0	2	EQ5DSelCar0	2
EQ5DPain0	5	EQ5DPain0	5	EQ5DPain0	5	EQ5DUusuAct18	0	EQ5DUusuAct0	2
EQ5DAnxDep0	5	EQ5DAnxDep0	5	EQ5DAnxDep0	5	EQ5DPain0	4	EQ5DPain0	4
EQ5DVAS0	4	EQ5DVAS0	4	EQ5DVAS0	4	EQ5DAnxDep0	4	EQ5DAnxDep0	4
ROL0	2	ROL0	2	ROL0	2	EQ5DVAS0	3	EQ5DVAS0	3
SOC0	2	SOC0	2	ROL18	0	ROL0	2	ROL0	3
PWB_overall0	8	SOC18	0	SOC0	2	SOC0	2	SOC0	3
PWB_overall18	0	PWB_overall0	8	PWB_overall0	8	PWB_overall0	8	PWB_overall0	9
SWB_overall0	4	SWB_overall0	4	SWB_overall0	4	SWB_overall0	4	SWB_overall0	5
EWB_overall0	1	EWB_overall0	1	EWB_overall0	1	EWB_overall0	1	EWB_overall0	2
FWB_overall0	1	FWB_overall0	1	FWB_overall0	1	FWB_overall0	1	FWB_overall0	2

selected hyperparameters for each model are presented in the Appendix Tables A4 and A5.

5.2.5 Model Evaluation

Models with the optimised hyperparameters were evaluated on test sets. The selected hyperparameters for each model are presented in Appendix Tables A4 and A5. Most of the performance metrics remained the same as in the previous study: accuracy, recall (also known as sensitivity), Area Under the Receiver Operating Characteristic Curve (AUC), as they are frequently used for evaluation in similar studies (Chapter 2). Furthermore, specificity was used as evaluation metric. It measures the proportion of correctly classified data points with negative label. It is very useful to include it in the evaluation, as together with recall, it provides the in depth analysis of model performance, accounted for model's tendencies to prioritise one of the classes. Since balanced accuracy, AUC, specificity and recall are sufficient metrics to understand the model performance and were the most common metrics in similar research, precision and F1-score were not considered in this chapter. Variable importance was also conducted on LR and RF models predicting 3 best performing outcomes at 18 weeks due to their explainability potential. Furthermore, ANOVA with Tukey HSD tests were performed to compare model performances. The methods for variable importance analysis (completed in Python) and significance testing (completed in R) were the same as in previous chapter.

5.2.6 Stakeholder Engagement

A consultation with three patient representatives was a part of a Data-driven Cancer Research Conference 2024 in Manchester, UK. As a part of the conference, patient representatives were allocated to presenters to enable discussions on research that are relevant to their experiences. The presented study included the design from the previous chapter, where the target variables were hospital utilisation outcomes. Patient representatives were asked about the challenges of going through the chemotherapy. They mentioned that HRQoL aspects were extremely difficult, as cancer and treatment affected their ability to perform daily activities. They stressed that everyone is aware of physical symptoms worsening during chemotherapy, but HRQoL aspects are often missed out of the conversation. They said that they would like to understand how chemotherapy would impact their every-day activities and social life, so that they could expect the changes in lifestyle. Consequently, these aspects were selected as target variables in this study.

The patients' perspective was further considered through the involvement of a clinical oncologist in all stages of the study, who also emphasised the importance of HRQoL changes during chemotherapy, which might be detrimental to patients' well-being. Furthermore, the conclusions from the Patient Centred Outcomes Research (PCOR) Group consultation for the design of the study reported in previous chapter were also taken into consideration. Therefore, the study includes variable importance analysis and ML model evaluation on original (unprocessed) data.

5.3 Results

Accuracy, recall, specificity and AUC of all models are provided in tables 5.4 and 5.5. The plots in Figure 5.2 represent AUCs of all target variable predictions apart from role functioning, due to its similarity to usual activities, but worse performance.

5.3.1 Physical Well-Being

Deterioration: Overall, the best performing model was LR with the highest AUCs at 6 (0.710), 12 (0.661) and 18 (0.761) weeks deterioration predictions. Even though LR did not provide the highest recall, it did not compromise the specificity unlike other models (e.g., MLP at 6 weeks: recall=0.902, specificity=0.162), preventing false negative predictions. All models had the best performance predicting deteriorations at 18 weeks, compared to 6 and 12 weeks.

Improvement: LR had the highest performance at 6 (AUC=0.718) and 18 weeks (AUC=0.662), and MLP had the highest performance at 12 weeks (AUC=0.843). MLP also had high recall (0.800) and specificity (0.886). Generally all models had poor recall (risking false improvement predictions), apart from LR (0.800) and aforementioned MLP (0.800) predicting improvement at 12 weeks. The models performed the best at 12 weeks, except for RF achieving highest AUC at 6 weeks.

5.3.2 Social Functioning

Deterioration: Overall, the best performing model was LR at 12 weeks (AUC=0.562) and at 18 weeks (AUC=0.685), and MLP at 6 weeks (AUC=0.595). The models were compromising either specificity or recall, with at least one of them being lower than 0.600. Overall, 18 weeks deterioration predictions had highest AUC for all models except MLP (highest AUC of 0.595 at 6 weeks).

Improvement: The best performing model was MLP at 6 (AUC=0.713), 12 (AUC=0.533) and 18 (AUC=0.736) weeks. These models also had highest recall, which was 0.750 at 6 weeks and 0.722 at 18 weeks with specificity being respectively 0.676 and 0.750. The MLP's recall at 12 weeks was only 0.200, suggesting a bad performance of this model. RF had the highest specificity for all time points, but compromised recall. All models achieved best predictive performance at 18 weeks.

5.3.3 Role Functioning

Deterioration: LR outperformed other models, achieving the highest AUCs at all time points: 6 (0.573), 12 (0.691) and 18 (0.675) weeks. Recalls of these models were higher than specificity, with the highest value at 12 weeks (0.750). The models had best overall performance at 18 weeks, even though LR was slightly better at 12 weeks.

Improvement: LR had the highest AUCs of all models at 6 (0.531) and 18 weeks (0.770), whilst MLP at 12 weeks (0.712). LR's performance at 18 weeks was the highest overall. This model also had the highest recall (0.688), but specificity of 0.852 was lower than XGB's specificity (0.902). Overall, models performed the

best at 18 weeks (except MLP with the highest AUC=0.712 at 12 weeks) and the worst at 6 weeks.

5.3.4 Usual Activities

Deterioration: The best performing model was LR at 6 (AUC=0.663) and 12 weeks (AUC=0.739), and MLP (AUC=0.718) at 18 weeks. These models also had highest recall, but not specificity, which was the highest for RF (6 and 12 weeks) and XGB (18 weeks). In general, specificity was a lot higher than recall. The models had the lowest performance at 6 weeks, while 12 and 18 weeks predictions had similar performance.

Improvement: LR outperformed all models at all time points with AUCs of 0.684 (6 weeks), 0.731 (12 weeks), and 0.898 (18 weeks). LR at 18 weeks was the highest performing model with excellent recall (0.900) and specificity (0.896). The specificity of all models was high, but often compromised recall. Models predicting 18 weeks improvement performed much better than models at other time points.

5.3.5 Mobility

Deterioration: The models had poor overall performance with MLP having the highest AUCs at 6 (0.529) and 12 (0.562) weeks, and LR (0.597) at 18 weeks. Recall of the models was poor as well, with the highest value of 0.560. Specificity values of the models were good overall, with the highest for RF at 6 weeks (0.973). It is unclear which time point resulted in the best predictions.

Improvement: MLP was the best performing model overall achieving AUC of 0.904 (recall=0.875, specificity=0.932) at 12 weeks. LR was the best models at 6 weeks (AUC=0.829) with very good recall (0.800) and specificity (0.859). At 18 weeks LR also outperformed other models with AUC of 0.715, high specificity (0.930), but low recall (0.500). Overall, 18 weeks predictions resulted in the lowest AUCs, while 6- and 12-week predictions depended on the model.

5.3.6 Significance Testing

ANOVA results suggest that at least one model's AUC mean is different from the rest ($F=4.202$, $p=0.007$) and Tukey HSD indicated that LR was significantly better than RF (0.011) and XGB (0.041). Outcome also affected prediction performance ($F=4.659$, $p=0.002$), as predictions of usual activities had higher mean AUC than role ($p=0.0363697$) and social ($p=0.001$) functioning. The time point of prediction was also a factor affecting AUC ($F=5.676$, $p=0.004$), with Tukey HSD indicating 18-weeks predictions resulting with significantly higher mean AUC than 6-weeks predictions ($p=0.003$). Welch Two Sample t-test indicated that improvement predictions were significantly better than deterioration predictions ($t=-3.079$, $df=101.39$, $p=0.003$).

Table 5.4: The results of HRQoL deterioration prediction at each time-point.

HRQoL outcome	Time-point of deterioration	Model	Accuracy	Specificity	Recall	AUC
Physical well-being	6 weeks	LR	0.716	0.676	0.745	0.710
		RF	0.625	0.324	0.843	0.584
		XGB	0.636	0.459	0.765	0.612
		MLP	0.591	0.162	0.902	0.532
	12 weeks	LR	0.663	0.656	0.667	0.661
		RF	0.613	0.438	0.729	0.583
		XGB	0.625	0.469	0.729	0.599
		MLP	0.638	0.656	0.625	0.641
	18 weeks	LR	0.766	0.727	0.795	0.761
		RF	0.675	0.424	0.864	0.644
		XGB	0.662	0.545	0.750	0.648
		MLP	0.740	0.667	0.795	0.731
Social functioning	6 weeks	LR	0.557	0.608	0.486	0.547
		RF	0.602	0.784	0.351	0.568
		XGB	0.591	0.765	0.351	0.558
		MLP	0.591	0.569	0.622	0.595
	12 weeks	LR	0.561	0.600	0.524	0.562
		RF	0.512	0.375	0.643	0.509
		XGB	0.463	0.425	0.500	0.463
		MLP	0.512	0.000	1.000	0.500
	18 weeks	LR	0.692	0.786	0.583	0.685
		RF	0.628	0.810	0.417	0.613
		XGB	0.564	0.738	0.361	0.550
		MLP	0.436	0.381	0.500	0.440
Role functioning	6 weeks	LR	0.568	0.540	0.605	0.573
		RF	0.580	0.820	0.263	0.542
		XGB	0.591	0.780	0.342	0.561
		MLP	0.432	0.000	1.000	0.500
	12 weeks	LR	0.695	0.632	0.750	0.691
		RF	0.646	0.646	0.841	0.631
		XGB	0.573	0.395	0.727	0.561
		MLP	0.524	0.395	0.636	0.516
	18 weeks	LR	0.675	0.674	0.676	0.675
		RF	0.675	0.837	0.471	0.654
		XGB	0.636	0.721	0.529	0.625
		MLP	0.662	0.628	0.706	0.667
Usual activities	6 weeks	LR	0.693	0.759	0.567	0.663
		RF	0.705	0.948	0.233	0.591
		XGB	0.670	0.828	0.367	0.597
		MLP	0.659	0.931	0.133	0.532
	12 weeks	LR	0.744	0.783	0.694	0.739
		RF	0.707	0.891	0.472	0.682
		XGB	0.671	0.826	0.472	0.649
		MLP	0.683	0.717	0.639	0.678
	18 weeks	LR	0.688	0.675	0.703	0.689
		RF	0.649	0.750	0.541	0.645
		XGB	0.688	0.800	0.568	0.684
		MLP	0.714	0.625	0.811	0.718
Mobility	6 weeks	LR	0.545	0.560	0.462	0.511
		RF	0.830	0.973	0.000	0.487
		XGB	0.716	0.827	0.077	0.452
		MLP	0.739	0.827	0.231	0.529
	12 weeks	LR	0.549	0.581	0.450	0.515
		RF	0.744	0.935	0.150	0.543
		XGB	0.683	0.806	0.300	0.553
		MLP	0.671	0.774	0.350	0.562
	18 weeks	LR	0.610	0.635	0.560	0.597
		RF	0.623	0.788	0.280	0.534
		XGB	0.662	0.808	0.360	0.584
		MLP	0.584	0.615	0.520	0.568

Table 5.5: The results of HRQoL improvement prediction at each time-point.

HRQoL outcome	Time-point of improvement	Model	Accuracy	Specificity	Recall	AUC
Physical well-being	6 weeks	LR	0.818	0.865	0.571	0.718
		RF	0.864	1.000	0.143	0.571
		XGB	0.807	0.932	0.143	0.538
		MLP	0.852	0.932	0.429	0.681
	12 weeks	LR	0.825	0.829	0.800	0.814
		RF	0.875	0.986	0.100	0.543
		XGB	0.888	0.957	0.400	0.679
		MLP	0.875	0.886	0.800	0.843
	18 weeks	LR	0.831	0.908	0.417	0.662
		RF	0.831	0.969	0.083	0.526
		XGB	0.844	0.954	0.250	0.602
		MLP	0.831	0.954	0.167	0.560
Social functioning	6 weeks	LR	0.591	0.632	0.450	0.541
		RF	0.750	0.926	0.150	0.538
		XGB	0.739	0.882	0.250	0.566
		MLP	0.693	0.676	0.750	0.713
	12 weeks	LR	0.671	0.776	0.200	0.488
		RF	0.817	1.000	0.000	0.500
		XGB	0.695	0.695	0.067	0.451
		MLP	0.744	0.866	0.200	0.533
	18 weeks	LR	0.679	0.683	0.667	0.675
		RF	0.833	0.983	0.333	0.658
		XGB	0.769	0.867	0.444	0.656
		MLP	0.744	0.750	0.722	0.736
Role functioning	6 weeks	LR	0.602	0.662	0.400	0.531
		RF	0.750	0.941	0.100	0.521
		XGB	0.659	0.824	0.100	0.462
		MLP	0.523	0.574	0.350	0.462
	12 weeks	LR	0.720	0.734	0.667	0.701
		RF	0.756	0.938	0.111	0.524
		XGB	0.720	0.875	0.167	0.521
		MLP	0.768	0.813	0.611	0.712
	18 weeks	LR	0.818	0.852	0.688	0.770
		RF	0.766	0.869	0.375	0.622
		XGB	0.792	0.902	0.375	0.638
		MLP	0.714	0.721	0.688	0.704
Usual activities	6 weeks	LR	0.761	0.797	0.571	0.684
		RF	0.807	0.905	0.286	0.596
		XGB	0.818	0.892	0.429	0.660
		MLP	0.773	0.838	0.429	0.633
	12 weeks	LR	0.829	0.861	0.600	0.731
		RF	0.854	0.958	0.100	0.529
		XGB	0.841	0.917	0.300	0.608
		MLP	0.829	0.889	0.400	0.644
	18 weeks	LR	0.896	0.896	0.900	0.898
		RF	0.909	0.985	0.400	0.693
		XGB	0.909	0.925	0.800	0.863
		MLP	0.857	0.851	0.900	0.875
Mobility	6 weeks	LR	0.852	0.859	0.800	0.829
		RF	0.875	0.949	0.300	0.624
		XGB	0.852	0.910	0.400	0.655
		MLP	0.830	0.846	0.700	0.773
	12 weeks	LR	0.866	0.905	0.500	0.703
		RF	0.902	0.919	0.750	0.834
		XGB	0.890	0.919	0.625	0.772
		MLP	0.927	0.932	0.875	0.904
	18 weeks	LR	0.896	0.930	0.500	0.715
		RF	0.896	0.958	0.167	0.562
		XGB	0.909	0.958	0.333	0.646
		MLP	0.883	0.915	0.500	0.708

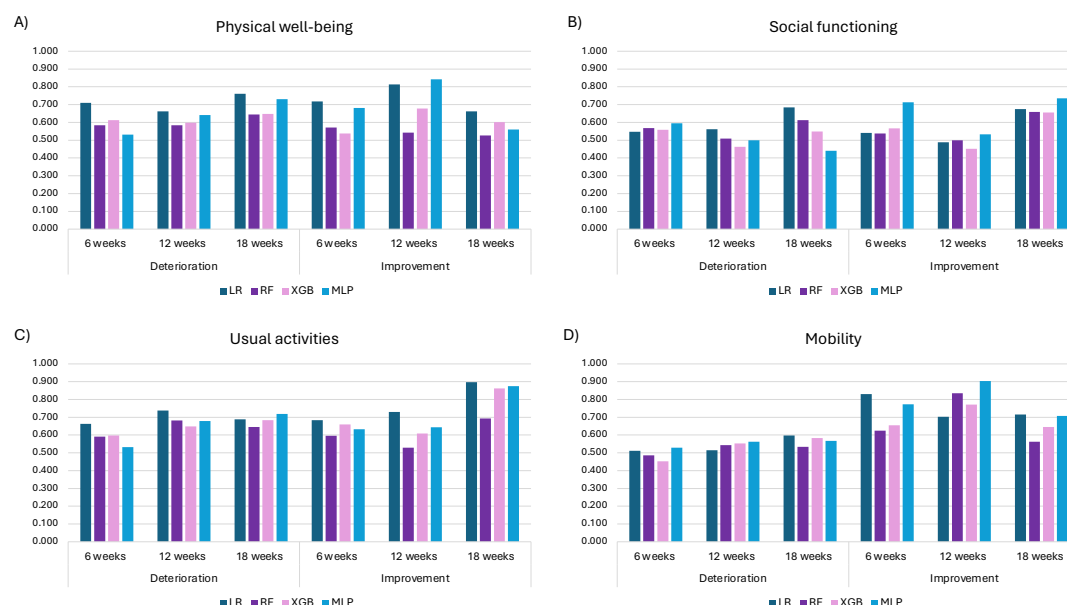


Figure 5.2: AUC values of models predicting A) physical well-being, B) social functioning, C) usual activities, D) mobility deterioration and improvement at 6, 12, and 18 weeks.

5.3.7 Variable Importance

Variables with coefficients for LR and values of importance for RF are presented in appendix tables A6,A7,A8. For prediction of physical well-being, LR looked mainly at clinical information (comorbidities and cancer characteristics). RF considered FACT-G and EQ-5D as most important predictors, as well as patients' BMI. LR in predicting mobility also found comorbidities the most meaningful predictors (including EQ5D mobility score at baseline for improvement prediction), while RF mainly looked at PROMs, cancer type and BMI. When predicting usual activities, LR also looked at the clinical characteristics and usual activities baseline score, whilst RF focused on BMI, cancer type and PROMs.

5.4 Discussion

5.4.1 Overall Results

HRQoL Changes During Chemotherapy

The predicted patient outcomes in this chapter were informed by the consultation with patient representatives, who highlighted the importance of understanding how chemotherapy could impact HRQoL outcomes. The findings suggest that ML models applied on PROMs, clinical and demographic data can successfully predict the fluctuations of HRQoL outcomes throughout chemotherapy. Models provided excellent performance in predicting improvement in physical well-being, usual activities and mobility at different stages of cancer treatment. These outcomes are particularly meaningful to patients, as they were selected following the discussion with patient representatives. According to the scoping review reported in Chapter 2, no previous

study reported patient engagement during the design of studies predicting patient outcomes using AI and PROMs. Therefore, the lack of focus on HRQoL outcomes in research has not been addressed before, even though it could inform treatment decisions [234]. Consultation with a clinical oncologist also endorsed the view that ML prediction of HRQoL changes during chemotherapy can be useful in clinical practice.

Model Performances

LR generally outperformed other models, which is a common outcome in medical research, as other models are more susceptible to overfitting [241]. In some cases, MLP achieved higher performance than LR. Nevertheless, the lack of explainability of MLP could affect public trust in this model. According to the consulted clinical oncologist, even difficult to interpret models should be considered useful, as long as they are used alongside other well-performing models, which enable explainability. This was provided through variable importance analysis. Similarly to the results in the study reported in Chapter 4, LR considered clinical information more important than PROMs, whilst RF mainly used PROMs for its predictions. Better performance of LR than RF adds an insight into understanding the predictive value of PROMs. Since RF was prioritising PROMs variables, ignoring clinical and demographic variables, PROMs predictive value might only support the model if combined with clinical and demographic information.

Improvement and Deterioration Predictions

Improvement predictions had generally higher performance than deterioration. This might be considered counterintuitive, as chemotherapy is usually associated with decline in HRQoL [236]. However, the deterioration might depend on patient characteristics from the start of chemotherapy. For example, metastatic disease may be more likely to show improvement due to higher burden of cancer symptoms prior to treatment. Therefore, the baseline values fed through the models potentially provided the information that is more useful for improvement predictions. Following the consultations with clinical oncologist, the predictions of positive changes in HRQoL measures can be equally useful.

Variable Importance

Similarly to the results presented in the previous chapter, LR tends to focus more on clinical data than on PROMs. However, RF is prioritising mainly PROMs. The PROMs variables which RF was considering important were from all 3 questionnaires. LR model frequently outperformed RF, especially in the deterioration predictions. This suggests that clinical variables are very important for prediction of outcomes. This further encourages the use of PROMs data as well as clinical and demographic data in model development. In this study, MLP has achieved a good overall performance. Unfortunately, its' explainability potential is limited. Therefore, it is important to consider other models as well to understand the balance between explainability potential and performance of the models.

5.4.2 Strengths and Limitations

Strengths

There are several strength of the research reported in this chapter:

1. The main strength of this work is the patient-centric approach achieved through the active engagement of a clinical oncologist and patient representatives throughout the study design. The representation of patients has enabled selections of outcomes that are truly important for patients, but rarely used in previous studies.
2. Including patient reports as input data in the models further considers patients' views. Three time points of HRQoL changes provided another insight into fluctuation of chemotherapy symptoms and understanding when individuals can be affected.
3. The rigorous reporting of data pre-processing methods, model development and evaluation supports the reproducibility of this study. The design followed the recommended methodology from the previous study.

Limitations

However, this work has its limitations:

1. The main limitation is associated with data collected in a clinical trial. While the ML models have been rigorously designed and evaluated, the data were subject to inclusion and exclusion criteria, which can lead to bias. Nevertheless, this chapter provides insights relevant to any AI research which predicts patient outcomes. It shows that including stakeholders can help understanding the relevant outcomes for patients, which would not have been predicted otherwise.
2. Temporal clinical trial data are usually relatively small samples and are affected by participants drop outs. This limits the ML methods that have been applied, yet the results are encouraging considering HRQoL aspects as outcomes.
3. Incomplete QLQ-C30 questionnaire used might have influenced the feature importance analysis and limited the exploration of a questionnaire that would be the most useful for predictions.

5.4.3 Conclusions

This chapter presented the work which successfully applied ML models on PROMs, clinical and demographic data to predict changes in HRQoL during chemotherapy, which could support preparation for adverse events and inform shared decision making about cancer treatments. The results further encourage the use of ML methods to identify factors related to chemotherapy toxicity and explore how cancer treatment affects individuals' lives.

The research presented here contributed to addressing two research questions:

1. **What is the added value of using PROMs in AI models predicting patient outcomes?** Even though Chapter 4 explored the predictive value of PROMs used as input data, this chapter focused on PROMs as outcomes. It has shown the potential of PROMs collected longitudinally to compute outcomes of deterioration and improvement of HRQoL measures during chemotherapy. However, the persistent importance of clinical and demographic variables suggests that PROMs do not provide enough information as a stand-alone input variables. Therefore, this chapter recommends utilising PROMs in addition to data that are gathered from patients.
2. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** This work added the perspective of patients on predicting outcomes during chemotherapy. While previous chapter only involved clinicians, this chapter gained important insights from patient representatives. This ensured that the predicted variables are meaningful for patients and clinically relevant. However, the stakeholder engagement in this study did not provide the comprehensive answer to this question. The stakeholder engagement should also involve more in-depth discussion on model development and feasibility of research. Therefore, these aspects will be explored further in this thesis.

The research reported in this chapter has further explored the use of PROMs collected at 1 time-point in ML models predicting patient outcomes. As mentioned earlier, these are the most common data types and models used in similar research. The following chapters will focus on longitudinal symptom severity reports and PROMs , and will build on the information acquired from this chapter and Chapter 4. Since hospital utilisation and patient HRQoL are clinically meaningful outcomes according to patients and clinicians, both outcomes will be predicted in the following chapters.

Chapter 6

Using ML and time-series PROMs To Predict Hospital Utilisation and HRQoL Changes

6.1 Introduction

So far, this thesis has focused on baseline PROMs collected at the start of the chemotherapy. Therefore, all input data in the models reported in Chapters 4 and 5 were collected at one time-point. This is the most common approach to predicting patient outcomes, as found in the scoping review (Chapter 2). In contrast, the scoping review showed that use of time-series PROMs was not common in previous studies. The methods that have been used in the limited research on time-series PROMs are mainly traditional ML models.

This chapter will start the exploration of time-series PROMs as predictors. It will contribute to addressing all research questions of the research reported in this thesis (listed in the order of relevance):

1. **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** This is the most important research question which this chapter is attempting to address, as it explores longitudinal PROMs as input data in ML models. These include both weekly symptom reports, and PROMs completed at the baseline, 6 and 12 weeks of the chemotherapy. Critical analysis of traditional ML models applied on temporal PROMs are conducted exploring the potential to predict patient outcomes. The models predicting hospital utilisation will be assessed focusing on long-term predictions (outcome happening at any time between when the report was completed and the end of study time-point at 18 weeks) and short-term predictions (outcome happening within 14-days from symptom report completion). A successful system that can alert to potential complications happening within a short time-frame or at any time during chemotherapy can be extremely useful to prevent serious complications or hospitalisation [70]. Understanding the models' ability to detect temporal patterns will provide novel insights into processing time-series PROMs.
2. **What is the added value of using PROMs in AI models predicting**

patient outcomes? Previous chapters indicated potential for PROMs data to improve the prediction models. Consequently, this chapter explores the added value of longitudinal PROMs for predicting patient outcomes, comparing with findings in the previous chapters. The performance of models which include and exclude PROMs data are compared to understand the added value of validated PROMs when symptom severity reports are available.

3. **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** This chapter builds on the knowledge acquired in the previous chapters, further strengthening the design of model development and evaluation. Chapter 4 has recommended using test set which is unaffected by any pre-processing methods. Therefore, data balancing was performed only on the train set. To further mitigate the bias which can arise from data leakage, nested Cross-Validation (CV) design is introduced in this chapter[242]. This design prevents the data leakage through all data pre-processing methods being done within CV loops.
4. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Patient representatives and a clinical oncologist were consulted on the research design, which helped understanding how to best process longitudinal patient-reported data. The outcomes (hospital utilisation and HRQoL variables) have been selected through patient and clinician involvement in Chapters 4 and 5. Furthermore, when the eRAPID clinical trial was designed, the choice of HRQoL outcomes was selected with extensive patient and clinician input.

As defined in Chapter 3 Section 3.3, symptom severity reports are patient-reported questionnaires used in eRAPID clinical trial, based on the established CTCAE criteria using patient-friendly language, collected longitudinally. They provide patients' perspective on their health status throughout the duration of treatment. The patterns in longitudinal reports can be identified by ML models to predict long-term, and short-term clinical outcomes [88]. However, the research on using time-series patient-reported data is still very limited, with majority of papers considering only PROMs collected at one time point (Chapter 2). The most common approach for processing longitudinally collected PROMs was considering temporality through coding it in the feature sets and using conventional ML models to predict outcomes. However, previous research has not focused on exploring the models' ability to capture temporal patterns in data within a set time (here 14 days) from symptom reporting, which can further utilise the longitudinality of reported PROMs in detecting sudden events. These short-term predictions have a potential to alert clinical staff and patients about potential deterioration, providing enough time to act preventatively. Exploring the added value of different types of longitudinal PROMs, i.e., symptom severity reports collected weekly and validated PROMs questionnaires collected at four time-points provides another novelty of this work and helps understanding the types of data which can be useful to include in ML models.

The outcomes predicted in this chapter are both hospital utilisation (admissions and triage events) and HRQoL changes during chemotherapy (the changes

from baseline to 18 weeks in mobility, physical well-being, social functioning, usual activities, and role functioning, measured by PROMs) in line with the methodology presented in Chapter 3. Hospital utilisation predictions can involve detecting changes in data happening over a long period of time, or sudden events. Therefore, this chapter will predict admissions and triage events within 14 days from the completion of a symptom severity report, and at anytime during chemotherapy (between completion of the report and the 18th week of the trial). The changes in HRQoL are computed for the 18th week of the chemotherapy, which is the end of study follow-up time-point for each patient in the eRAPID clinical trial [70].

6.2 Methods

6.2.1 Methodology for Using Machine Learning to Predict Outcomes from Longitudinal PROMs

Following the main aim of this thesis, the research in this chapter is conducted to predict both types of patient outcomes during chemotherapy: hospital utilisation and HRQoL changes. Logistic regression (LR), random forest (RF), multilayer perceptron (MLP), and extreme gradient boosting (XGBoost) are developed. The models have been selected due to their common use in similar studies [230]. Additionally to the comparison of the ML models' performances in predicting outcomes, this study will assess the added value of using PROMs when longitudinal symptom severity reports are available, as well as evaluate robustness of 2 different data pre-processing pipelines (nested CV vs grid search CV) A clinical oncologist was involved in the study design to ensure relevance of this study to clinical practice.

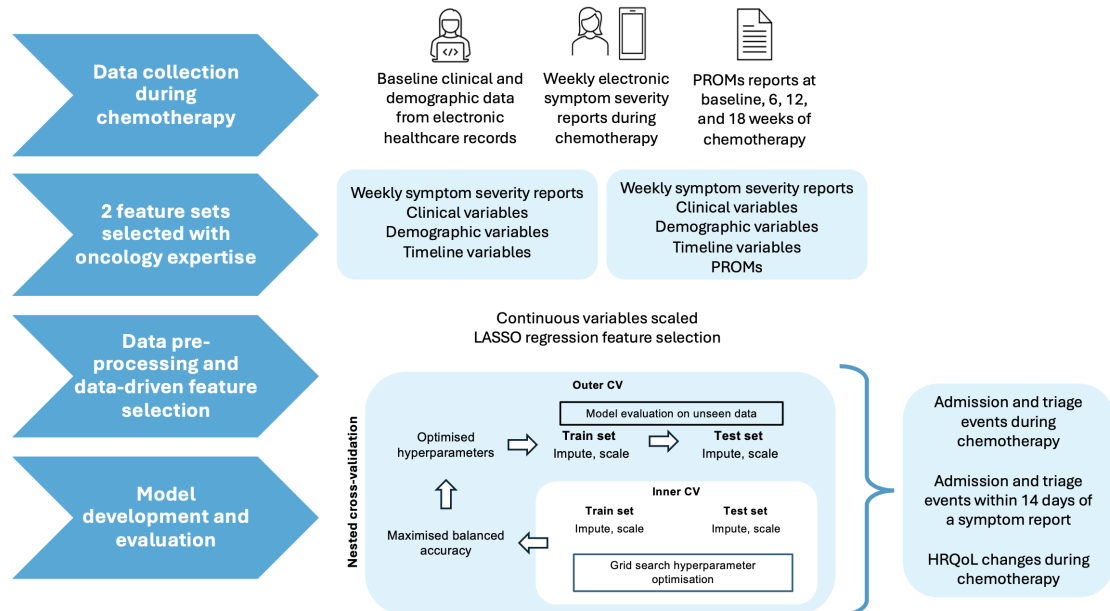


Figure 6.1: Flow diagram illustrating methodology in this chapter. Four models were applied on data collected throughout eRAPID clinical trial to predict outcomes during chemotherapy. Nested cross-validation was used to develop and evaluate models. Models built on data including and excluding PROMs were compared.

6.2.2 Dataset

Input Variables

The dataset used in this chapter has been described in Chapter 3, Section 3.4. All input variables will be used in the development of all models predicting all target variables. Models with and without PROMs data are trained to assess the added value of using validated PROMs when symptom severity reports are available for the model.

Target Variables

Hospital utilisation during chemotherapy: The following hospital utilisation target variables are considered in this chapter:

- Admission (whether a patient was admitted to hospital at any time between when the report was completed and the end of study time-point at 18 weeks)
- Triage (whether a patient contacted the acute oncology ward hotline at any time between when the report was completed and the end of study time-point at 18 weeks)
- Admission within 14 days (whether a patient was admitted to hospital within 14 days from completing the report)
- Triage within 14 days (whether a patient contacted hospital within 14 days from completing the report)

Admission and triage events were also chosen as outcomes in Chapter 4, but no short-term predictions were considered before. The 14-day window of admission was chosen by a clinical oncologist as a useful time frame for detecting symptom deterioration in clinical practice, which would enable preparing for or avoiding potential admission. The 14-day window for triage is equally useful, as it would help prepare for increased hospital burden and potential treatment complications. The 14-day time-frame is estimated based on symptom development and the time it may take for patients to address the concerns. This time-frame is based on the oncologists' expertise and feedback from patients who participated in the eRAPID clinical trial. The variable indicating whether or not the outcome happened within the 14 days from a symptom report was computed from the dates of admissions and triage episodes during the clinical trial and the time-steps of symptom severity reports. A binary column was created to indicate if the event happened (1) or not (0), following the report. Changes to chemotherapy is the only outcome that is not added in this chapter, even though it was an outcome in Chapter 4. This is due to the lack of meaningful temporal data on the times and reasons for changes in treatment.

HRQoL changes during chemotherapy: The outcomes in this chapter follow the changes in HRQoL measures selected through consultation with patients for predictions reported in Chapter 5. This chapter focuses solely on deterioration, excluding improvement from the analysis, as predicting deterioration offers more actionable insights. Therefore the deterioration of the following self-reported outcomes were considered in this chapter:

- Mobility (whether patient's self-reported mobility deteriorated during chemotherapy), measured with EQ-5D
- Physical well-being (whether patient's self-reported physical well-being deteriorated during chemotherapy), measured with FACT-G.
- Social functioning (whether patient's self-reported social functioning deteriorated during chemotherapy), measured with EORTC QLQ-C30.
- Role functioning (whether patient's self-reported role functioning deteriorated during chemotherapy), measured with EORTC QLQ-C30.
- Usual activities (whether patient's self-reported ability to perform usual activities deteriorated during chemotherapy), measured with EQ-5D.

The deterioration of each HRQoL aspect was calculated from the change from baseline PROMs to 18-weeks PROMs and thresholds, using MCID [237, 238], as described in Chapter 5 Section 5.2.2 and Table 5.1. The HRQoL outcomes are only considered as long-term outcomes (the deterioration within 14-days from the symptom report is not considered), as the collection of PROMs was at the same time-points for each patient. Therefore, short-term outcome would not be possible to measure.

6.2.3 Data Pre-processing

Variable Preparation

The selection of variables was conducted via consultation with a clinical oncologist in order to ensure clinical relevance. Only symptoms (collected through symptom severity reports) which were common for all cancer types were considered (pain, nausea, vomiting, diarrhoea, temperature, chills, physical ability, appetite, fatigue, sore mouth, indigestion, shortness of breath). Scores for reaction of injection site, phlebitis, sore hands, neuropathy, taste, nail changes, nose bleeds, skin rash and ascites have not been used, as their completion depended on cancer type.

Furthermore, aside from the total number of all comorbidities, a binary variable indicating whether any of respiratory comorbidities exist (chronic obstructive pulmonary disease, emphysema, asthma, chronic bronchitis) was created. A variable indicating a number of days from previous chemotherapy cycle was also created due to the importance of the cycle stage in symptom development.

Variable selection was further conducted using Least Absolute Shrinkage and Selection Operator (LASSO) regression with cross-validated minimised lambda value to improve model performance and interpretability, automatically select relevant features, and prevent overfitting [243]. This chapter improves the variable selection from previous chapters (where manual removal of correlated variables has been conducted) by using LASSO to enable a quick, automated and robust way to reduce dimensionality and perform feature selection.

Since symptom severity reports and PROMs have been collected at different time-points, PROMs variables are selected to match the time-steps of the reports.

The PROMs value was allocated to each report depending on the time-point of completion. For example, for each report completed between baseline and 6 weeks, the PROMs collected at baseline were allocated; for reports between 6 and 12 weeks, the PROMs collected at 6 weeks were allocated. Repeated values of patient's static data (clinical and demographic variables) are added to each report. These steps were necessary, as all data types were collected at different time-points. PROMs, clinical and demographic data had to be matched to symptom severity reports to ensure the input is suitable for ML models to ensure the input is suitable for ML models.

Class imbalance

The class imbalance in hospital utilisation target variables is more prominent than in the HRQoL changes (Table 4.4). The positive class was always smaller than the negative class for admissions and triage variables, whilst the imbalance in the HRQoL outcomes varied depending on variables (Table 6.1). To balance classes in train sets, a technique called SMOTE was used. This technique generates new unique sample, unlike the random sampling with replacement technique used in (Chapter 4 and 5) [244]. Therefore, instead of using random oversampling to obtain a quick baseline model (as in the previous chapters), this chapter uses SMOTE for oversampling of the minority class. Balancing classes is performed to help a model learn from the data which are originally imbalanced. While random oversampling enables quick prototyping, SMOTE can be more effective as it has lower risk of overfitting by avoiding exact duplication and improves model generalisation by creating more representative training data for the minority class [245].

Table 6.1: Class distribution in each hospital utilisation and HRQoL outcome

Hospital utilisation		HRQoL changes	
Variable	Class distribution	Variable	Class distribution
Admission	no: n=2620 (80%) yes: n=640 (20%)	Physical well-being	no: n=163 (44%) yes: n=219 (56%)
Admission within 14 days	no: n=3085 (95%) yes: n=175 (5%)	Social functioning	no: n=209 (54%) yes: n=179 (46%)
Triage	no: n=2048 (63%) yes: n=1212 (37%)	Role functioning	no: n=214 (56%) yes: n=171 (44%)
Triage within 14 days	no: n=2806 (86%) yes: n=454 (14%)	Usual activities	no: n=201 (52%) yes: n=183 (48%)
		Mobility	no: n=260 (92%) yes: n=22 (8%)

Addressing irregularity in data

There are three different aspects introducing irregularity in the dataset:

1. **Irregular reporting** - The irregularity of reporting varied between patients, which resulted in many missed or too frequent symptom severity reports. This

irregularity is often closely related to patients' disease status. Therefore, patient representatives were consulted during study design and asked about potential reasons for missing reports or reporting more often than asked. The patient representatives suggested that reasons for the irregularity might inform healthcare status (for example, missing reports due to feeling unwell, or feeling better and not perceiving the benefit of reporting). Therefore, a variable which indicates the time interval between the current and previous report has been added to the data. Consequently, models are given the information on irregularity in the dataset and can learn from it.

2. **Missing data within variables** - There were very few cases of missing data points in data resulted from incomplete PROMs completion or incomplete information on clinical and demographic data. The number of missing values per column is provided in the Supplementary File Table A9. No symptom severity report has any missing data, as patients did not have an option to skip a question. Any unfinished reports have been removed from analysis. K-Nearest Neighbours (KNN), ($k=5$)[239] has been used to impute any sporadic missing data within completed reports, PROMs, clinical or demographic data.
3. **Outliers** - The number of reports completed by three patients exceeded the expected number (eighteen reports) by over three times. The histogram (Chapter 3, Figure 3.3) also clearly suggests that these patients were outliers. The reports of these patients were removed, as these patients were identified as outliers. The longest sequence after removing the data of the outliers was 43 reports per person.

6.2.4 Model Development

The pipeline for model development has changed from the Chapters 4 and 5. In the previous chapters, the cross-validation was performed only using grid search for hyperparameter optimisation. In this chapter, the methodology for model development is improved through using nested CV. This method can ensure more robust hyperparameter optimisation, prevent data leakage and over-estimation of results, the models were implemented through the loops of nested CV [246] (Figure 6.1). This pipeline has a potential to improve the robustness of model development as all stages of data pre-processing are completed separately within the nested CV loops. The hyperparameter tuning was performed in the inner CV loop and in outer CV loop model was evaluated on unseen test set. In each loop a stratified split of patients to train and test sets was completed. In each of the loops scaling of continuous variables, missing data imputation, and SMOTE have been performed, as described in previous section. LR, RF, MLP, and XGBoost were developed with grid search hyperparameter tuning, performed on train set. The nested CV included 5 iterations of both inner and outer loops.

6.2.5 Model Evaluation

The models are evaluated with five different performance metrics. Balanced Accuracy (BA) was selected as the main metric for both hyperparameter optimisation and model evaluation, as it is useful at assessing performance of models serving as

a screening system, being able to capture model's bias towards one class. Including BA is another improvement from the previous chapter as it can lead to more robust model evaluation, unaffected by class imbalance in test sets [247]. Area under the ROC curve (AUC) was also used, as in previous chapters, due to being very common metric in ML studies. Even though it is not recommended for imbalanced data, it can be used for between-studies comparison. Recall, specificity, and accuracy are also computed, as in previous chapter.

ANOVA [248] with Tukey HSD tests were performed to compare model performances. The variable importance analysis is performed on LR and RF models predicting admission during chemotherapy, admission within 14 days from the self-reported symptom severity report, physical well-being and usual activities.

6.2.6 Stakeholder Engagement

The involvement of clinical oncologist throughout the design of research reported in this chapter was extremely valuable for input variable selection. The data contained many comorbidities, and the relevant ones were selected with the clinician, as described in Section 6.2.3. Furthermore, the time window of 14 days from the symptom report for the short-term predictions or admission and triage was chosen following the discussion with the clinical oncologist.

Three patient representatives from the PCOR group were interviewed individually to support with the trial design. They voluntarily offered to take part in the interviews following the email to research advisory group of the PCOR team informing about this opportunity. The interviews consisted of a brief presentation introducing the general concepts, and questions to patient representatives regarding the potential reasons of the irregularities in reporting symptoms, views on the outcome variables, and the overall opinions on the design. The discussion provided valuable insights into potential reasons for skipping reports or reporting more frequently than requested. It also provided further insights into the usefulness of HRQoL outcomes predictions, in addition to hospital utilisation. Patient representatives appreciated the investigation of the outcomes affecting their daily activities, such as usual activities and role functioning. Furthermore, the involvement of clinical oncologist helped selecting relevant input data and choosing the time-frame of short-term hospital utilisation predictions.

6.3 Results

The results of the hospital utilisation and HRQoL changes predictions are presented in Table 6.2. These are the results of the models trained with all available variables (selected through LASSO regression), and developed with nested CV pipeline (Figure 6.1). The model performance is compared against the average of highest reported AUC (mean=0.791, median=0.784) values in all papers which used time-series data to predict patient outcomes, analysed in scoping review (Chapter 2).

Table 6.2: Results of all models predicting hospital utilisation and HRQoL changes outcomes.

Evaluation metric	Model	Hospital utilisation				HRQoL changes				
		Admission	Triage	Admission within 14 days	Triage within 14 days	Mobility	Physical well-being	Social functioning	Usual activities	Role functioning
Balanced accuracy	LR	0.956	0.782	0.640	0.848	0.919	0.694	0.755	0.855	0.770
	RF	0.967	0.970	0.598	0.702	0.986	0.997	0.995	0.976	0.993
	MLP	0.955	0.915	0.648	0.722	0.905	0.908	0.938	0.896	0.933
	XGBoost	0.989	0.836	0.580	0.532	0.965	0.864	0.878	0.920	0.959
AUC	LR	0.988	0.876	0.668	0.902	0.970	0.748	0.827	0.879	0.845
	RF	0.999	0.997	0.848	0.928	0.999	1.000	1.000	1.000	1.000
	MLP	0.988	0.971	0.729	0.837	0.992	0.963	0.976	0.986	0.994
	XGBoost	1.000	0.917	0.821	0.687	0.987	0.941	0.961	0.980	0.992
Recall	LR	0.944	0.722	0.577	0.934	0.929	0.698	0.757	0.924	0.766
	RF	0.938	0.955	0.206	0.460	1.000	1.000	0.995	1.000	0.994
	MLP	0.952	0.904	0.417	0.592	0.995	0.928	0.920	0.990	0.912
	XGBoost	0.983	0.795	0.171	0.085	0.934	0.824	0.856	0.947	0.950
Specificity	LR	0.969	0.842	0.703	0.762	0.910	0.691	0.753	0.831	0.774
	RF	0.997	0.985	0.990	0.944	0.973	0.995	0.995	0.953	0.992
	MLP	0.958	0.925	0.878	0.852	0.815	0.889	0.956	0.802	0.953
	XGBoost	0.995	0.877	0.988	0.979	0.996	0.904	0.901	0.894	0.969
Accuracy	LR	0.964	0.797	0.696	0.786	0.927	0.695	0.755	0.874	0.770
	RF	0.985	0.974	0.948	0.877	0.998	0.998	0.995	0.995	0.992
	MLP	0.957	0.917	0.854	0.816	0.980	0.911	0.939	0.970	0.936
	XGBoost	0.992	0.912	0.944	0.853	0.992	0.912	0.944	0.853	0.961

6.3.1 Hospital Utilisation

Comparison of long-term and short-term predictions

Overall, models predicting admission and triage at any time during chemotherapy have better performance than models predicting admissions and triage within 14 days from a completed report. Long-term predictions achieved all evaluation metrics of over 0.9 for admissions (the highest BA=0.989 for XGBoost, the lowest BA=0.955 for MLP), suggesting very well-balanced performance. All metrics are over 0.7 for triage (the highest BA=0.970 for RF, the lowest BA=0.782 for LR). With AUC's of both long-term predictions reaching 0.9, which is higher than average performance of similar papers (0.791), the models can be considered to have good overall performance.

The performance of short-term predictions is much worse than long-term predictions. Even through the AUC values reach 0.8 for admissions (the highest BA=0.648 for MLP, the lowest BA=0.580 for XGBoost) and 0.9 for triage (the highest BA=0.848 for LR, the lowest BA=0.532 for XGBoost) within 14 days from the completed report, the other metrics reveal that the models are struggling with higher class imbalance. As presented in Figure 6.2 C), the models in most cases had much higher specificity than recall, which means that the models were usually biased towards negative class. BA captures this better than AUC, as shown in Figure 6.2 A) and B). This suggests that ML models do not predict outcomes limited in short time-frame well from the longitudinal data, which is important for addressing patients' needs. The short-term predictions will be further investigated in the next Chapter (7), using DL approach.

Comparison of admission and triage predictions

Overall, all models achieved better performance predicting admission at any time during chemotherapy than triage at any time during chemotherapy, as shown in Table 6.2. Both admission and triage predictions have improved, compared to results reported in Chapter 4. This suggests that adding temporal data have positively impacted predictions. Short-term predictions of triage events tend to have higher

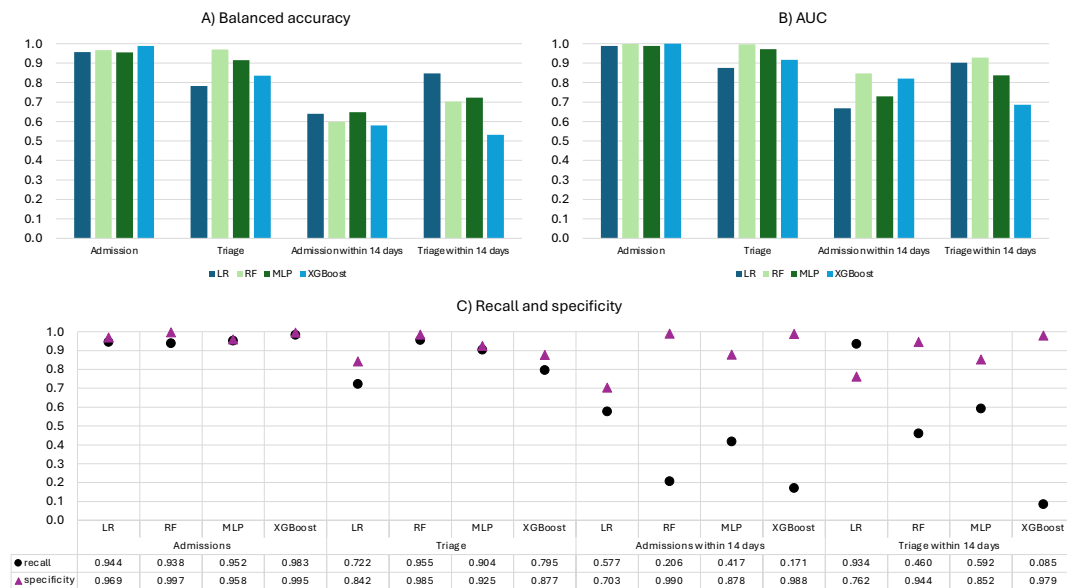


Figure 6.2: Balanced accuracy, AUC, specificity and recall plots for hospital utilisation outcomes predicted by all models.

value of performance metrics than short-term predictions of admissions for all models, apart from XGBoost. As illustrated in Figure 6.2, these differences are not big for all of the models, apart from LR, which achieved BA of 0.848 predicting triage within 14 days from the report, and 0.640 predicting admission within 14 days from the report.

Comparison of model performances

Based on AUC values, RF was the best performing model for all hospital utilisation outcomes. It also had the highest BA for triage during chemotherapy predictions (0.970). For triage within 14 days from the completed report prediction, LR achieved the highest BA of 0.848, which is noticeably better than the performance of other models. For short-term (BA between 0.580 and 0.648) and long-term (BA between 0.955 and 0.989) admissions predictions, the performance did not vary much between models, suggesting that all models have a similar ability to predict these outcomes (Figure 6.2 A)).

Added value of PROMs

Models were trained with and without PROMs data added to longitudinal symptom severity reports, clinical and demographic data. This is to further explore the predictive value of PROMs when symptom severity reports are available. As shown in Table 6.3, PROMs have always improved the performance of models, which further endorse the view that PROMs add predictive value to models predicting hospital ad-

Table 6.3: Comparison of results of models trained on data including and excluding PROMS predicting admission and admission within 14 days.

Evaluation metric	Model	Admission		Admission within 14 days	
		Including PROMs	Excluding PROMs	Including PROMs	Excluding PROMs
Balanced accuracy	LR	0.956	0.679	0.640	0.595
	RF	0.967	0.885	0.598	0.562
	MLP	0.955	0.859	0.648	0.637
	XGBoost	0.989	0.911	0.580	0.564
AUC	LR	0.988	0.746	0.668	0.632
	RF	0.999	0.987	0.848	0.839
	MLP	0.988	0.942	0.729	0.744
	XGBoost	1.000	0.988	0.821	0.795
Recall	LR	0.944	0.683	0.577	0.503
	RF	0.938	0.781	0.206	0.131
	MLP	0.952	0.820	0.417	0.349
	XGBoost	0.983	0.834	0.171	0.137
Specificity	LR	0.969	0.675	0.703	0.688
	RF	0.997	0.988	0.990	0.993
	MLP	0.958	0.898	0.878	0.925
	XGBoost	0.995	0.987	0.988	0.992
Accuracy	LR	0.964	0.677	0.696	0.678
	RF	0.985	0.947	0.948	0.947
	MLP	0.957	0.883	0.854	0.894
	XGBoost	0.992	0.957	0.944	0.946

mission outcomes. The difference between models including and excluding PROMs was bigger for models achieving higher performance (admission outcome), rather than lower (admission within 14 days).

6.3.2 HRQoL Changes

Comparison of HRQoL outcomes

Overall, models predicting the deterioration in HRQoL outcomes achieved a very well-balanced performance, which at least one model reaching 0.9 for all performance metrics computed. Mobility had the best overall performance, as all models achieved very high values of all metrics. The worst overall performance was achieved by models predicting physical well-being, with the lowest performance of LR (BA=0.694), as shown in Figure 6.3 A). However, the performance of RF predicting physical well-being was still much higher (BA= 0.997).

Comparison of model performances

The model which achieved the best performance in predicting all HRQoL target variables is RF (Figure 6.3 A)). Not only had it AUC and BA of over 0.9 for each outcome, but also both specificity and recall, suggesting a more balanced performance than of other models (Figure 6.3 C)). LR had the lowest performance for all outcomes but mobility, where MLP had slightly worse BA. Specificity and recall for LR were usually lowest compared to other models, achieving similar values (Figure

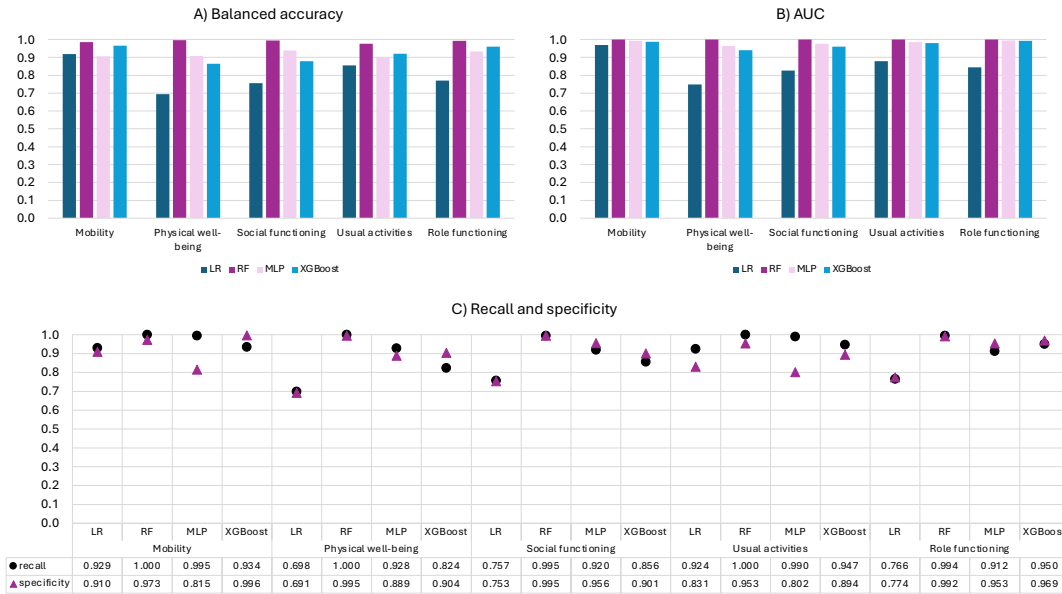


Figure 6.3: Balanced accuracy, AUC, specificity and recall plots for HRQoL outcomes predicted by all models.

6.3 C)). MLP and XGBoost had varied performances, which were depending on the outcome.

6.3.3 Significance Testing

Hospital utilisation outcomes

In predicting hospital utilisation outcomes, ANOVA analysis did not indicate any significant differences between model performances ($F=0.426$, $P=0.735$). However, there was a significant difference between the performance of models predicting different outcomes ($F=12.32$, $P < .001$). The follow-up Tukey HSD indicated that the performances of models predicting admission within 14 days were significantly worse than of models predicting admission ($P < .001$), and that models predicting triage within 14 days were significantly worse than of models predicting triage ($P < .05$). There was no statistically significant difference of short-term triage and admission outcomes, and long-term triage and admission outcomes.

HRQoL outcomes

In predicting HRQoL deterioration, ANOVA results suggest that there is a significant difference in model performance ($F=46.14$, $P < .001$). Tukey HSD indicated that the performances of RF, MLP and XGBoost were statistically significantly higher ($P < .001$) than LR. RF was significantly better than MLP, and XGBoost than RF, with the significance level of $P < .01$. ANOVA analysis also indicated

that there is a significant difference between predicted outcomes, but the level of significance is lower ($F=2.539$, $P < .05$). Tukey HSD suggested that this difference was between mobility and physical well-being only ($P < .05$).

6.3.4 Variable Importance

The variables and their importance values are listed in Appendix Table A10 for short- and long-term admission predictions and in Appendix Table A11 for HRQoL deterioration predictions. LR considered similar variables as the most important in short- and long-term admission predictions. Vomiting and stoma were the top two important variables, followed by some clinical information, such as previous chemotherapy (for short-term predictions) and diabetes (for long-term predictions). The variables considered important by RF differed between these outcomes. For short-term predictions, self-reported symptoms of nausea, pain, appetite, fatigue and indigestion were the most important. For long-term prediction, the top five variables were BMI, week of completion, age, emotional well-being (FACT-G) and nausea.

For predictions of physical well-being, LR considered mainly clinical variables including cancer stage (primary or metastatic), reported stoma, respiratory co-morbidities and diabetes plus vomiting (the only self-reported symptom). However, RF focused on physical well-being at baseline, age, BMI, disease site and employment status. For usual activities predictions, stoma, sex, disease site and two EQ-5D variables: self-care and usual activities at baseline were the most important variables considered by LR. RF focused mainly on usual activities at baseline, BMI, physical ability, age and fatigue.

6.4 Discussion

6.4.1 Overall Results

The results reported in this chapter suggest that ML models trained on longitudinally collected symptom severity reports, PROMs, clinical and demographic data can successfully predict hospital utilisation and HRQoL changes happening at any time during chemotherapy following a symptom report. The models achieved AUC higher than average and amongst highest of similar papers analysed in scoping review (Chapter 2). The performances were also well balanced (not skewed to any class), which shows the even distribution of sensitivity and recall, as well as BA of over 0.9.

The predictions of admission happening within 14 days from a completed report were not reliable, as the models tended to misclassify positive cases (admission and triage events). However, short-term predictions of outcomes during chemotherapy can be extremely useful, as they can alert medical staff and patients of potential complications in advance, providing enough time to act to prevent deteriorations [70]. Therefore, short-term predictions will be further explored in the next chapter with the use of more complex modelling approach than ML.

Overall, RF is the best performing model in this chapter for both hospital utilisation and HRQoL outcomes, compared to other models. This is consistent with other studies analysed in a systematic review of time-series classification techniques used in biomedical applications [249]. RF is an ensemble method, which means that it combines output of multiple ML models. This process can help preventing over-fitting and manage high-dimensional data well [250], which might explain RF outperforming other ML models predicting outcomes from time-series PROMs. This performance of long-term predictions can be considered very good, comparing to other studies, and looking at all evaluation metrics being around 0.9. However, the performance of RF predicting short-term outcomes should be improved.

Long-term admission and triage predictions exceeded the performance of hospital utilisation predictions of models based on only baseline data from the research reported in Chapter 4. The models predicting HRQoL outcomes also outperformed models trained only on baseline data (Chapter 5). Therefore, this chapter further encourages the collection and use of symptom severity reports in ML research predicting patient outcomes. Using longitudinal patient-reported data reveals the patients' perspective on their health and increases the patients' participation in symptom management [251].

The potential explanation for poor performance of predicting admissions within 14-day interval could be the inherent class imbalance, which was increased due to the shorter time-frame of interest. Furthermore, this performance could be explained by exploring differences between classes. Table 6.4 presents the number of variables which are statistically significantly different for classes in each outcome. It indicates that more features differed significantly for classes in admission and triage than for classes in admission and triage happening within 14 days. These differences might have provided a more structured discriminatory advantage for ML models in predicting long-term hospital utilisation outcomes than short-term (when the time-frame is limited).

Chapter 4, as well as previous studies [6] have already established the predictive value of PROMs, for their use in ML models predicting patient outcomes and emphasised the high value of patients' subjective view on their health status. This chapter further explored the value of PROMs in ML models when symptom severity reports are available. The results suggest that PROMs improved model performance, and were among the most important features influencing the predictions. For example, emotional well-being collected with FACT-G questionnaire is an important predictor for RF predicting long-term hospital admission, and self-care collected through EQ-5D is amongst top predictors of usual activities for RF. These results further encourage the collection and use of standardised PROMs questionnaires, even when other patient-reported data are available. However, it also stresses the importance of including clinical and demographic variables in models. For example, having stoma and respiratory comorbidities are two of the top predictors of both hospital admissions and HRQoL changes for LR.

Table 6.4: Number of features within different significance levels of the difference between classes for each outcome.

Hospital utilisation outcomes				
	Admission within 14 days	Triage within 14 days	Admission	Triage
p<0.001	14	16	20	21
p<0.01	6	6	8	8
p<0.05	5	6	1	4
p>0.05	18	15	14	10

HRQoL outcomes					
	Mobility	Physical well-being	Social functioning	Usual activities	Role functioning
p<0.001	32	28	34	34	33
p<0.01	2	1	3	3	5
p<0.05	4	2	1	1	1
p>0.05	5	12	5	5	4

6.4.2 Strengths and Limitations

Strengths

The following factors are the strengths of the work reported in this chapter:

1. The main strength of this chapter is using longitudinally collected patient-reported data. This was extremely rare in the literature, as reported in Chapter 2. As the results show, utilising the temporal patterns in data for predicting patient outcomes can improve model performance.
2. Utilising the variety of patient-reported data should also be considered the strength of this chapter. Including standardised PROMs as well as symptom severity reports with different patterns of longitudinal reporting can be challenging, but provides a more comprehensive representation of patients' views on their health status.
3. Another strength of this chapter is the prediction of an outcome that happens within the short time frame of patient-reported symptom severity. The 14-day window from a symptom report to hospital admission or triage can be crucial for preventing deterioration, as advised by a consulted clinical oncologist. If applied in clinical practice, such system could prevent hospital admission or patients deterioration.
4. Furthermore, this chapter involved two main changes in methodology:
 - (a) The models implemented through nested CV. This method is known to have a potential to reduce bias in model selection and prevent data leakage [242, 246]. This method can be therefore considered an improvement from the standard grid search method used previously, as all data handling is

performed within the CV loops. This protects the test set further from being affected by data pre-processing, and the train set from receiving any information from test set.

- (b) Balancing data with SMOTE. This method can be more effective than random oversampling, as it decreases the risk of overfitting and improving model generalisation by creating more representative data [245].

Limitations

Apart from the inherent bias only of using clinical trial data, affecting the representability of population (discussed in Chapter 4 Section 4.4.2 and Chapter 5 Section 5.4.2, the research reported in this chapter has the following limitations:

1. ML models might not be able to process the complexity of time-series data. Longitudinal reports provide an opportunity for more complex modelling using DL methods, such as recurrent neural networks [88]. These could improve the short-term admissions and triage predictions following the reports.
2. ML models require one type of input data, so all values from static variables were repeated for each symptom report. This enabled combining static with longitudinal data, but might have created noise.

Both of these limitations will be addressed in the following chapter. All further limitations that are relevant for the whole thesis will be discussed in Chapter 8.

6.4.3 Conclusions

The research reported in this chapter critically analysed the application of traditional ML models to clinical, demographic and longitudinal patient reported data (PROMs and symptom severity reports) to predict hospital utilisation and HRQoL deterioration during chemotherapy. These results encourage collecting time-series PROMs and utilising them in AI research.

This chapter contributed to addressing the research questions in this thesis:

1. **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** The research reported here compared the performance of four different ML models. The good overall performance of RF was highlighted, but the potential of more complex DL models will be further explored in the next chapter.
2. **What is the added value of using PROMs in AI models predicting patient outcomes?** This chapter explored the added value of longitudinal PROMs and symptom severity reports. The results suggest better performance in predicting all outcomes of models in this chapter, compared to models built on static data only (Chapter 4 and 5). Furthermore, the research reported here suggest that PROMs improve model performance even when symptom severity reports are available.

3. **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** The results obtained through models developed in a nested CV pipeline were compared with the performance of models built with simple grid search only (as in Chapters 4 and 5). Nested CV led to better results overall. This method has a potential to prevent bias and data leakage, and therefore is recommended for using PROMs in AI research.
4. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Patient representatives and clinical oncologist were involved throughout the design of the research reported in this chapter. This has helped strengthening the methodology through selecting clinically relevant variables and pre-processing data. The outcomes predicted by the models in this chapter were be both hospital utilisation and HRQoL variables, which are following patients' and clinicians' recommendations.

The following chapter will further explore the predictions of hospital utilisation and HRQoL changes throughout chemotherapy from longitudinal PROMs and symptom severity reports. It will adopt a DL model suitable to process time-series data to assess its potential at processing complex temporal patterns, and improve the performance of predicting admissions and triage events within 14 days from a symptom report. Patient involvement will further support the design of Chapter 7.

Chapter 7

Using LSTM and time-series PROMs To Predict Hospital Utilisation and HRQoL Changes

7.1 Introduction

The previous chapter showed that validated PROMs can improve model performance, even when longitudinal symptom reports are available and already included in the model. The time-series data were used to predict outcomes happening at any time between when the report was completed and the end of study time-point at 18 weeks (long-term predictions) and within 14 days from the completed report (short-term predictions). The results suggested that ! (!)ML models can achieve much better performance of long-term predictions than short-term predictions. Even though long-term predictions are valuable for personalised patient support, the short-term prediction can provide more preventative opportunities, alerting clinical staff and patients on potential deterioration timely.

Chapter 6 used only ML models to analyse time-series data. This is also a common way to analyse longitudinal data in previous studies, using either statistical models, or classical ML methods [212, 213, 223, 252]. These methods have limited capability to capture complex dependencies and patterns in multi-dimensional or time-series data [186]. Therefore, this chapter completes addressing two following research questions:

1. **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** This chapter presents the research applying Long-Short Term Memory (LSTM) model to clinical, demographic, and longitudinal patient-reported data to predict hospital utilisation and Health-Related Quality of Life (HRQoL) during chemotherapy. The results are compared with the ML models applied in Chapter 6. The comparison helps understanding if Deep Learning (DL) models can be more effective when analysing longitudinal data for both long- and short-term predictions.
2. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Even though the previous chapters had a component of stakeholder engagement to understand how to process time-

series data, this chapter further uses patient representatives' views when applying DL models to data. This is important to consider, as LSTM model require equal sequences. However, the reports completed in the eRAPID were irregular (patients not completing all required reports at expected time-frames).

There are existing studies exploring DL models for predictions of oncology outcomes, but the main focus is on toxicity symptoms [88, 208, 235], rather than hospital utilisation. These studies have used LSTM model, which is the most common approach to process time-series data in healthcare [186].

Out of the explored studies using DL methods for oncology outcome predictions, only Wang et al., in a paper predicting head and neck cancer symptoms [88], and its follow-up study [208] used patient-reported data. Similarly to this research, their work involved longitudinally collected symptom reports. Both of these studies describe a common challenge of using longitudinally collected patient-reported data, which is the irregularity in patients' reports arising from skipping questionnaires completion. Therefore, they explored missing data imputation techniques which handle the irregularity. Nevertheless, these techniques might introduce bias and prevent the model from learning from the patterns of missingness [253], which could hold information on patients' health status. Furthermore, Wang et al., [208] have only used patient-reported data without the inclusion of clinical or demographic information. This chapter explores the novel approach of combining all these types of data collected at various irregular time-points.

Combining time-series data with static data (e.g., demographic) forms an additional challenge in DL models. A simple approach is a repetition of the same static information for each time-point generated by the same patient [186]. However, there are possibilities to use the DL network architecture to handle static data inclusion in a more efficient way. Time-series and static data can be processed by different network layers as separate inputs. This is the most common approach in medical outcome predictions using static and temporal data [186]. For example, Li et al., [254] used a multi-modal fusion approach that uses the representation of static data as the initial hidden state of RNN. Another study put static input separately through a fully connected layer, and time-series input through LSTM and convolutional neural network layers, combining both outputs at the end [255].

Although such studies predict clinical outcomes, there is also no evidence of patients or clinicians engagement during the study design. Since the development of AI models is not based on patients' perspective or clinical knowledge, it might not actually serve its purpose or address needs of diverse group of people in an equal and fair way [256]. Involving patients and public is essential not only for the feasibility of novel models, but also for increasing public trust in AI research [37]. This issue is true to all AI research which aim to predict patient outcomes. However, when using time-series PROMs, patient representatives can provide valuable insights into the patterns of reporting and help shaping the data pre-processing steps.

According to the literature review (Chapter 2), there is no existing research predicting chemotherapy toxicity outcomes using DL models and time-series patient-reported data combined with static clinical and demographic information.

Chapter 6 indicated that classical ML models can process time-series data to predict admissions and triage during chemotherapy with a good performance. However, these methods struggled with the short-term predictions of these outcomes. This is a key limitation, as the timely predictions can support prevention of deterioration and hospital utilisation. Therefore, the proposed approach uses the LSTM model to predict admissions and triage events within 14 days from the completion of a symptom severity report. Long-term hospital utilisation outcomes and HRQoL changes are also predicted. Furthermore, in this study the information from irregularity of reporting was used by providing the model with a variable indicating the time between reports. The missing reports were not imputed. Therefore, the model has an opportunity to learn from the irregularities in reporting. This approach was decided following the consultations with patient representatives and a clinical oncologist during the study design.

7.2 Methods

7.2.1 Methodology for Using LSTM to Predict Outcomes from Time-Series PROMs

The research reported in this chapter compares the performance of LSTM (implemented in this chapter) and traditional ML models: XGBoost, RF, LR, and MLP (implemented in the previous chapter), using clinical, demographic, and longitudinally collected patient-reported data to predict hospital utilisation and HRQoL changes during chemotherapy. All models are developed using nested Cross-Validation (CV) pipeline, as recommended in Chapter 6. Patient representatives and a clinical oncologists were consulted during the design of this research. The overview of the methodology is presented in Figure 7.1.

7.2.2 Dataset

The input data used in this chapter include exactly the same variables as the data used in Chapter 6, described in Section 6.2.2, and derived from all available, presented in Table 3.2 in Chapter 3. The target variables are also the same as in previous chapter (Section 6.2.2), including the calculations of MCID threshold for HRQoL changes (Table 5.1).

7.2.3 Data Pre-Processing

Variable Preparation

The PROMs value has been allocated to each report depending on the time-step of completion, due to the varied time-steps throughout the clinical trial. The static data (clinical and demographic variables) have been kept separately from time-series data (symptom severity reports and PROMs) during the data pre-processing for LSTM. As described in Chapter 6, for the traditional ML models, the static data have been repeated for each time-step, as these models cannot take separate inputs.

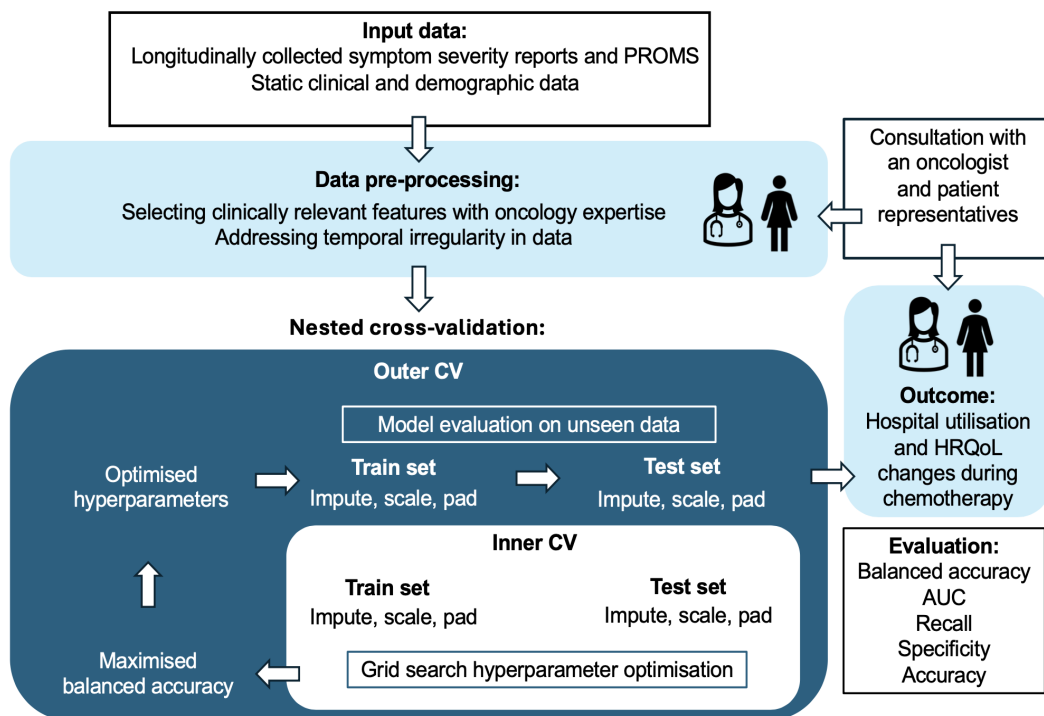


Figure 7.1: Flow diagram illustrating data flow and the nested cross-validation (CV) pipeline for implementing LSTM. The methodology for implementing ML models was described in Chapter 6.

Addressing Irregularity in Data (Missing Reports)

LSTM model requires all sequences of data to have the same length for each subject. However, the number of reports submitted by each patient varied due to missed or too frequent reporting. The challenge of irregularity can be addressed in two different ways: 1) by aggregating the reports that were completed more than once a week (e.g., taking average or the lowest score out of all reports completed during that week) and imputing the missing reports [88], or 2) by leaving the irregularity in data, and including a variable which indicates the time interval between the current and previous report. The longest sequence is then selected and all sequences are padded (filled with a particular value) to match its length. This is a common approach to adjust the size of the data in DL approaches [257]. Due to a significant variability in frequency of reporting, illustrated in Chapter 3 Figure 3.3, option 1) would create too much synthetic data and would prevent the model from learning on real data. Therefore, a variable has been created to indicate the number of days since the previous report. It provides the opportunity for the model to learn from the temporal pattern of reporting.

According to patient representatives, missing reports or reporting too often could hold information on how each patient is feeling. Consequently, padding sequences has been performed within the nested cross-validation loops. Missing reports were not imputed, but k-nearest neighbours (KNN, $k=5$)[239] has been used to impute any sporadic missing data and the reports of 3 patients who were identified as outliers have been removed from the dataset, as in Chapter 6.

7.2.4 Model Development

LSTM model architecture

The model accepts 2 separate inputs: static data (demographic and clinical information) and time-series data (symptom severity reports and PROMs). The LSTM layer processes time-series data, and outputs the hidden representation of time-series sequences. The fully connected layer receives the concatenated input of static data and the LSTM output (processed temporal dependencies). The output is converted by sigmoid activation function to a probability for binary classification. Therefore, the final output is 0 or 1 indicating the prediction of event not happening (0) or happening (1). The model architecture is presented in Figure 7.2. The hidden size, number of layers, learning rate were established through hyperparameter tuning. Due to large class imbalance in data (Table 6.1), class weight factor is set to 1. This is a common approach in studies using deep learning architecture to prevent creating synthetic data [258].

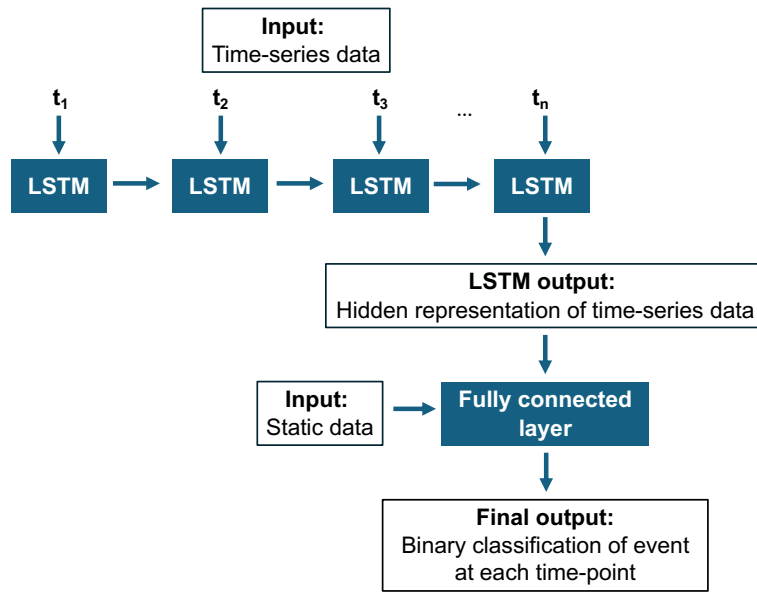


Figure 7.2: Flow diagram illustrating the LSTM model. Time-series input is processed through LSTM layer, which output is then passed through fully connected layer, receiving separate static input. The model hyperparameters were established during nested cross-validation process.

Nested cross-validation

To ensure robust hyperparameter optimisation, prevent data leakage and over-estimation of results, the models have been implemented through the loops of

nested cross-validation [246] (Figure 7.1). The hyperparameter tuning has been performed in the inner cross-validation loop and in outer cross-validation loop model was evaluated on unseen test set. In each loop a stratified split of patients to train and test sets is completed. The continuous variables are scaled to unit variance, and missing values are imputed with KNN(k=5). The sequences were then padded to 43 time-steps (the longest sequence for a patient after removing outliers). The padded value is -999, in order to avoid confusion, as the dataset does not include the value of -999 originally. To further avoid model learning from padded data, masking has been applied on padded values. All these steps are performed within loops to avoid data leakage.

The hyperparameter tuning in the inner loop involves the following hyperparameter sets: hidden size: [64, 128], number of layers: [1, 2], and learning rate:[0.001, 0.0001]. Therefore, there are 8 combinations of hyperparameters. The inner loop uses 2 folds, with each iteration running for up to 20 epochs. In the outer loop, the model with the best hyperparameters from tuning is trained for 50 epochs and evaluated based on the average performance metrics across 3 iterations. The number of epochs is based on literature and experimental observation of performance. The inner loop is responsible mainly for hyperparameter tuning so 20 epochs is computationally efficient and sufficient the purpose of optimisation. However, once the hyperparameters are selected, training the model for 50 epochs provides the model with more training opportunities, ensuring more robust performance estimate [259]. Early stopping is applied if the validation loss does not improve for 5 consecutive epochs, which can prevent model overfitting [259]. Training uses the BCEWithLogitsLoss function, which combines a sigmoid activation with binary cross-entropy for binary classification, due to the required binary output (event happening or not happening) [260].

The pipeline for all traditional ML is similar, with some adjustments to ensure model compatibility. In each loop, the training set is balanced using Synthetic Minority Oversampling Technique (SMOTE) to help the model learn from the data which are originally imbalanced. This step was unnecessary in LSTM, as the class weight factor was adjusted. Furthermore, ML models do not require as much running time as LSTM, so the nested cross-validation included 5 iterations of both inner and outer loops. In LSTM, to support code efficiency, the number of iterations is limited [206]. Another difference between the two models is combining static and temporal data. Since ML models cannot take two separate inputs, the static data have been repeated for each time-step.

7.2.5 Model Evaluation

The models are evaluated with 5 different performance metrics. As in Chapter 6 (Section 6.2.5), balanced accuracy is selected as the main metric, as it is useful at assessing performance of models serving as a screening system, being able to capture model's bias towards one class. AUC, specificity, recall and accuracy are also reported for in-depth model understanding. Variable importance has also been performed on the LSTM model predicting usual activities, physical well-being, admission at any time during chemotherapy, and admission within 14 days from the

report. The same outcomes are explained by variable importance analysis in Chapter 6. However, due to no in-build explainability function of the model, Shapley Additive Explanations (SHAP) [261] values are computed to aid interpretability of the model. This is a common method for variable importance analysis in similar studies [262]. To assess the differences between model performances, as in previous chapters, ANOVA has been performed, followed by Tukey HSD tests.

7.2.6 Stakeholder Engagement

As in all previous stages, a clinical oncologist has involved throughout the design of this research. Patient representatives from PCOR group were also consulted for research reported in this chapter, together with Chapter 6. It has helped understanding that imputing missing reports and aggregating frequent ones might prevent the model from learning the information that is hidden in the timelines of the reports. It was extremely important to consider, as LSTM only processes input with the same sequence length.

No thematic analysis were performed after the interviews, as there were only three patient representatives interviewed and the discussions were not recorded. These interviews were used exclusively to guide the design of the research reported in this chapter.

7.3 Results

This section presents the comparison of LSTM models and ML models developed in Chapter 6 developed to predict all hospital utilisation and HRQoL changes variables. The results of all models are presented in Table 7.1.

7.3.1 Hospital Utilisation

Overall, the predictions of hospital utilisation have a similar performance across all outcomes. Balance accuracy and AUC for all outcomes are above 0.7, which suggests that the performance is no better than average of AUC values (0.791) reported by papers analysed in scoping review (Chapter 2). The specificity values are slightly higher than recall in all cases, suggesting that the model's tendency towards the majority class was mitigated.

In comparison with ML models, LSTM has not improved the long short-term hospital utilisation predictions (Table 7.1). All ML models achieved a better performance than LSTM predicting both admission and triage. However, LSTM (BA = 0.780) has outperformed ML models in predicting admission within 14 days from a completed report. This indicates that the complexity of time-series pattern detection for the short-term prediction is more effective using LSTM. However, this has not been the case for the triage within 14 days predictions, where LR (BA=0.848) and MLP (BA=0.722) perform better than LSTM (BA=0.706). In both short-term outcome prediction the recall and specificity have the lowest difference in LSTM than in other models, suggesting an ability to provide balanced performance, even with large class imbalance (as shown in Chapter 6, Table 6.1).

Table 7.1: Results of LSTM models developed in this chapter and ML models developed in chapter 6 predicting all outcomes of hospital utilisation and HRQoL changes.

		Hospital utilisation				HRQoL changes				
		Admission	Triage	Admission within 14 days	Triage within 14 days	Mobility	Physical well-being	Social functioning	Usual activities	Role functioning
Balanced accuracy	LR	0.956	0.782	0.640	0.848	0.919	0.694	0.755	0.855	0.770
	RF	0.967	0.970	0.598	0.702	0.986	0.997	0.995	0.976	0.993
	MLP	0.955	0.915	0.648	0.722	0.905	0.908	0.938	0.896	0.933
	XGBoost	0.989	0.836	0.580	0.532	0.965	0.864	0.878	0.920	0.959
	LSTM	0.717	0.776	0.780	0.706	0.672	0.464	0.562	0.620	0.901
AUC	LR	0.988	0.876	0.668	0.902	0.970	0.748	0.827	0.879	0.845
	RF	0.999	0.997	0.848	0.928	0.999	1.000	1.000	1.000	1.000
	MLP	0.988	0.971	0.729	0.837	0.992	0.963	0.976	0.986	0.994
	XGBoost	1.000	0.917	0.821	0.687	0.987	0.941	0.961	0.980	0.992
	LSTM	0.775	0.839	0.845	0.779	0.683	0.478	0.562	0.715	0.944
Recall	LR	0.944	0.722	0.577	0.934	0.929	0.698	0.757	0.924	0.766
	RF	0.938	0.955	0.206	0.460	1.000	1.000	0.995	1.000	0.994
	MLP	0.952	0.904	0.417	0.592	0.995	0.928	0.920	0.990	0.912
	XGBoost	0.983	0.795	0.171	0.085	0.934	0.824	0.856	0.947	0.950
	LSTM	0.618	0.823	0.715	0.699	0.918	0.586	0.710	0.830	0.872
Specificity	LR	0.969	0.842	0.703	0.762	0.910	0.691	0.753	0.831	0.774
	RF	0.997	0.985	0.990	0.944	0.973	0.995	0.995	0.953	0.992
	MLP	0.958	0.925	0.878	0.852	0.815	0.889	0.956	0.802	0.953
	XGBoost	0.995	0.877	0.988	0.979	0.996	0.904	0.901	0.894	0.969
	LSTM	0.816	0.729	0.846	0.712	0.425	0.342	0.414	0.410	0.930
Accuracy	LR	0.964	0.797	0.696	0.786	0.927	0.695	0.755	0.874	0.770
	RF	0.985	0.974	0.948	0.877	0.998	0.998	0.995	0.995	0.992
	MLP	0.957	0.917	0.854	0.816	0.980	0.911	0.939	0.970	0.936
	XGBoost	0.992	0.912	0.944	0.853	0.992	0.912	0.944	0.853	0.961
	LSTM	0.773	0.765	0.836	0.709	0.875	0.482	0.552	0.781	0.905

Best performing ML models achieved balanced accuracy and AUC of over 0.9 in long-term predictions of hospital utilisation, which is higher than average in similar studies and suggests good overall performance of the models.

7.3.2 HRQoL Changes

The predictions of HRQoL changes using LSTM are poor for all outcomes (frequently indicating models' ability to classify outcomes only slightly better than guessing), apart from role functioning (BA=0.901). The predictions of mobility, physical well-being, social functioning, and usual activities has not achieved BA of over 0.7. This suggests that the model is not well balanced.

Even though the performance of LSTM (BA = 0.901) in predicting role functioning is higher than average in similar studies, it performed worse than most of the ML models (RF, MLP and XGBoost). This difference was large, with the best model (XGBoost) achieving BA of 0.959. LSTM only outperformed LR (BA = 0.770) predicting this outcome. Nevertheless, LSTM predicting other HRQoL outcomes has much worse performance than all ML models developed in Chapter 6. These differences are presented in Figure 7.3. The differences between recall and specificity of LSTM predicting HRQoL outcomes are much bigger than for ML models (Table 7.1). This suggests that the LSTM model has a tendency to bias towards one class, which is opposite to predicting hospital utilisation.

7.3.3 Significance Testing

The results of Tukey HSD test do not suggest any significant difference for any model predicting long- and short-term hospital utilisation outcomes. This means

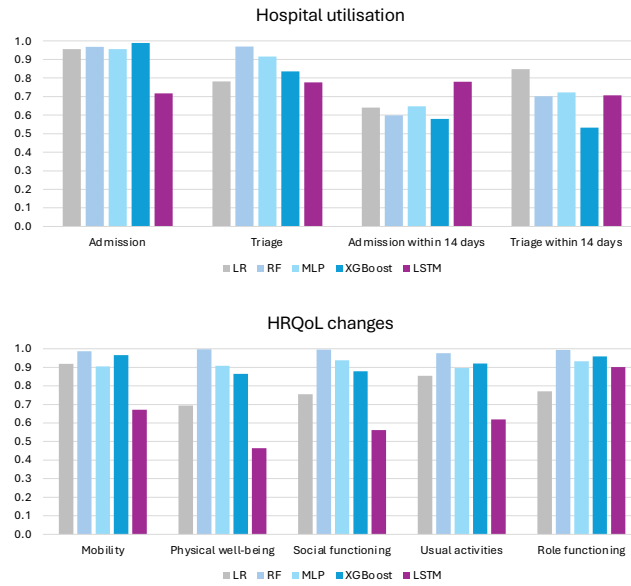


Figure 7.3: Balanced accuracy values for LSTM and ML predictions of all outcomes (plotted on the right side).

that the advantage of using LSTM for admissions and triage within 14 days from the report over other ML models is not statistically significant.

Tukey HSD analysis revealed the differences between LSTM model and ML models predicting hospital utilisation and HRQoL outcomes. LSTM model performs statistically significantly worse than LR, RF, XGBoost and MLP ($P < .001$). This level of the significance in the differences suggests that simpler ML models might be more suitable than LSTM for predicting outcomes.

7.3.4 Variable Importance

The SHAP values, which attempt to explain variable importance of LSTM models, are plotted for short- and long-term admission predictions in Figure 7.4. SHAP values computed for features predicting admission indicate that the top five predictors are role functioning, global well-being, days since previous report, social functioning and physical well-being. The low values of role functioning, global well-being and days since previous report have high impact on being admitted to hospital, and high values of social functioning and physical well-being have a high impact on not being admitted to hospital. The top five predictors of admission within 14 days are role functioning, social functioning, functional well-being, days since chemotherapy cycle and age. Role and social functioning are the most significant predictors, but the direction of the impact is not as apparent as in the long-term admission predictions. Functional well-being and

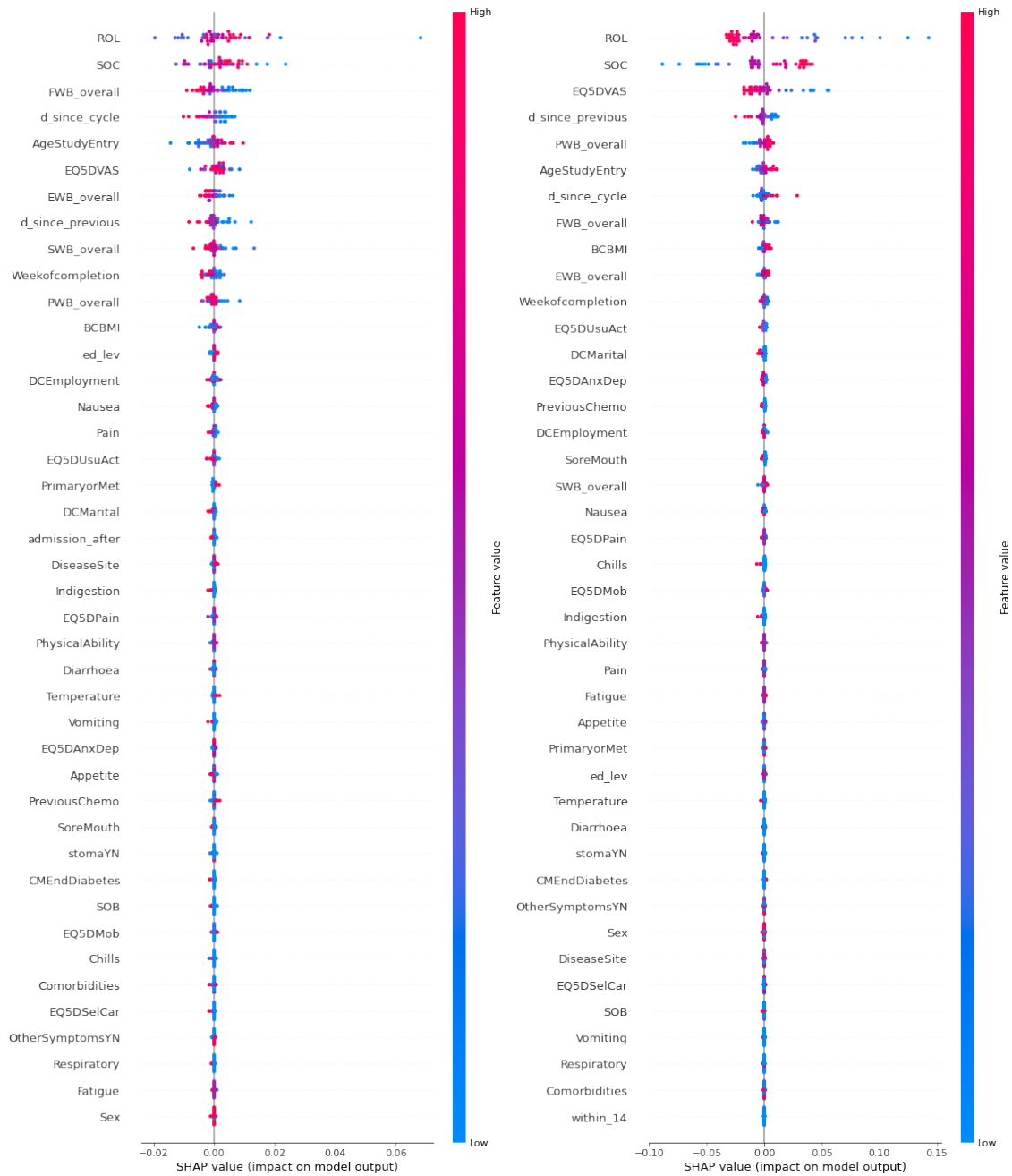


Figure 7.4: SHAP plots indicating feature importance for short- and long-term admission predictions. ROL - role functioning, SOC - social functioning, FWB_overall - functional well-being, d_since_cycle - days since last chemotherapy cycle, EWB_overall - emotional well-being, d_since_previous - days since previous report, SWB_overall - social well-being, ed_lev - level of education, EQ5DUUsAct - usual activities, PrimaryorMet - primary or metastatic, DCmarital - marital status, EQ5DPain - pain, EQ5DAnxDep - anxiety and depression, PreviousChemo - previous chemotherapy (yes,no), SOB - shortness of breath, EQ5DMob - mobility, EQ5DSelCar - self care

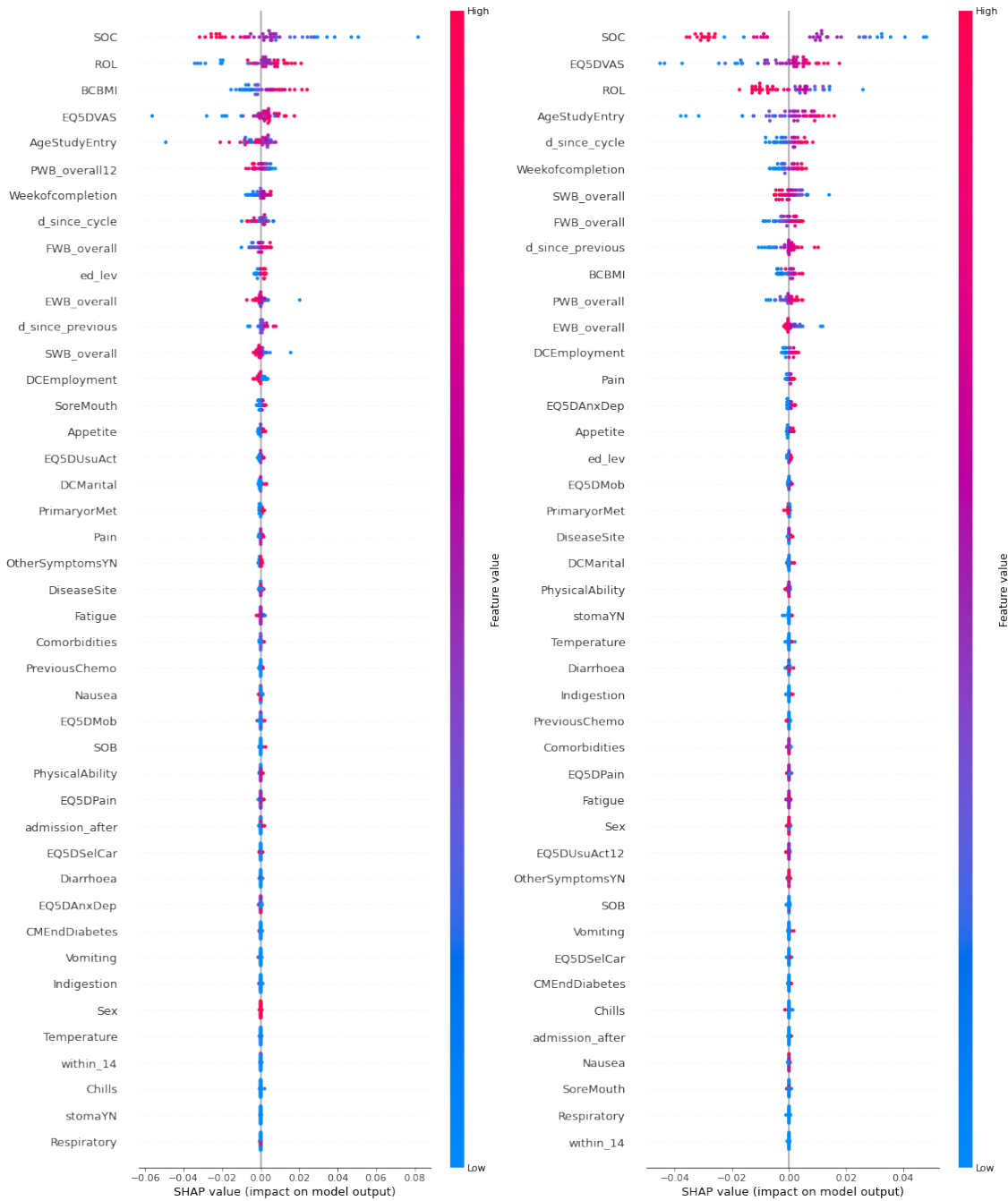


Figure 7.5: SHAP plots indicating feature importance for physical well-being and usual activities predictions. ROL - role functioning, SOC - social functioning, FWB_overall - functional well-being, d_since_cycle - days since last chemotherapy cycle, EWB_overall - emotional well-being, d_since_previous - days since previous report, SWB_overall - social well-being, ed_lev - level of education, EQ5DUuAct - usual activities, PrimaryorMet - primary or metastatic, DCmarital - marital status, EQ5DPain - pain, EQ5DAnxDep - anxiety and depression, PreviousChemo - previous chemotherapy (yes,no), SOB - shortness of breath, EQ5DMob - mobility, EQ5DSelCar - self care

days since chemotherapy cycle are also important, with low values predicting positive class. Older age is also positively associated with being admitted in 14 days from the report. Weekly symptom reports do not appear to be important variables in predicting hospital utilisation, contrary to the clinical expectations.

The SHAP values for physical well-being and usual activities predictions were computed to represent the variable importance for HRQoL changes and are plotted in Figure 7.5.2. The top five predictors of physical well-being are social functioning, role functioning, BMI, global well-being and age. Low values of social functioning, and high values of role functioning, BMI, and global well-being are associated with deterioration in physical well-being. Age is also an important predictor of well-being with no obvious direction of impact. The top predictors of usual activities are social functioning, role functioning, global well-being, age and days since chemotherapy cycle. In usual activities predictions, low values of social functioning and role functioning, and high values of global well-being, age, and days since chemotherapy cycle were the most important predictors of deterioration. It is important to notice that the LSTM models' predictions of HRQoL were poor, so the variable importance information might not be as valuable as for hospital utilisation.

7.4 Discussion

7.4.1 Overall Results

The results of this chapter presented the comparison of LSTM model and ML models developed in Chapter 6 to predict hospital utilisation and HRQoL changes during chemotherapy from longitudinal symptom reports, PROMs, clinical and demographic data. ML models outperformed LSTM in predictions of all HRQoL changes, and long-term admission and triage predictions. However, LSTM performance was better for predicting admission within 14 days from a completed report. In short-term triage predictions, the performance of LSTM is similar to simpler ML models, but recall and specificity showed a better ability of the model to predict positive and negative class with similar performance, without skewing predictions towards one class.

The overall better performance of LSTM than ML in short-term hospital predictions, especially admissions, is consistent with literature suggesting that DL models are able to process more complex and sudden fluctuations in the dataset [263]. Therefore, in time-series scenarios where the outcomes happen shortly after the input time-step, deep network architectures, such as LSTM, have more potential to succeed. The main issue with short-term prediction performance of ML models was generating a lot of false negative results.

On the other hand, the predictions of admission and triage which happened following the report at any time during chemotherapy were better for ML than LSTM. There is previous evidence that ML can perform better than DL. For example, XGBoost often outperforms DL models when predicting outcomes from numerical, tabular data [206]. Predicting outcomes happening at any time during chemotherapy can require less complex processing of longitudinal data than

outcomes within a short time-frame (here 14 days). This complexity was removed in the long-term predictions scenario, which might have strengthened the performance of ML, without causing much difference for LSTM.

The overall performance of models predicting hospital utilisation is good, which can be compared with other studies. Peterson et al., [223], predicted emergency department visits and hospital admissions during chemotherapy from clinical, demographic and patient-reported data using ML models. The reported AUC for the best model was 0.783, which is lower than the highest AUC of LSTM for short-term admission predictions (0.845), and XGBoost for admission during chemotherapy predictions (0.989). In another study predicting emergency visits and hospital admissions during cancer treatment [212], the AUC of gradient tree boosting was 0.798, which is also lower, compared to the results in this study. The good performance of these models might be the result of robust design with nested cross-validation used for hyperparameter tuning, as this approach can reduce bias in model selection [242].

The HRQoL deterioration predictions were much better for ML models than for LSTM. Low specificity of LSTM models in HRQoL predictions suggests their bias towards the positive class. This means that the model predicted deterioration in the HRQoL better than improvement or no change. Even though this bias can be problematic for most screening models, in this case, detecting deterioration in HRQoL can be more useful in clinical practice than detecting improvement or no change. HRQoL outcomes were predicted only for patients who were adhering to the clinical trial without dropping-out until the 18th week. Therefore, the models predicting HRQoL outcomes were applied to smaller datasets than the models predicting hospital utilisation. The low performance of LSTM was potentially the result of the dataset size, as DL models require larger and complex datasets than simpler ML approaches [230]. The difference in dataset size did not affect ML performance. The differences between the majority of features for the classes in all outcomes were statistically significant (Table 6.4, Chapter 6), which provided discriminatory patterns in the features, captured by ML.

The LSTM model predicting late symptoms and patient-reported outcomes for head and neck cancer patients in the study of Wang et al., [208] was much more successful (AUC between 0.8 and 0.9) than LSTM model in the current study predicting HRQoL outcomes. The difference might have resulted from different techniques for addressing irregularity in reports. Consulting patient representatives has shown that the patterns in irregular reporting might hold information about the health status and well-being. Therefore, this chapter focused on development of LSTM model which can learn from the irregularity in data, even though it inherently requires equal sequences for every patient. In contrast, Wang et al., generated a lot of synthetic data using data imputation methods to address irregularity in data [208]. This approach can reduce representativeness of a given sample and may lead to biased estimate and invalid conclusions [264].

7.4.2 Strengths and Limitations

Strengths

The research reported in this chapter has notable strengths:

1. The robust design of model development and evaluation pipeline is the main strength of this chapter. The model implementation with the nested cross-validation pipeline enabled robust hyperparameter tuning and minimised the risk of data leakage [246]. Additionally, multiple evaluation metrics reported enabled in-depth explanation of model performance and between-study comparison. This allowed critical analysis of model performance and challenging some of the findings in earlier research.
2. The comparison of LSTM and simpler ML models predicting long- and short-term outcomes during chemotherapy provided in-depth analysis of the potential of these models to detect temporal patterns in data. LSTM is considered a state-of-the-art model for time-series data analysis. However, the results in this chapter suggest that such complex models might not be necessary when simpler models can process the information more efficiently.
3. Using SHAP to attempt the interpretation of LSTM is another strength of the research reported in this chapter. This allowed further investigation into which variables are important for predictions, and the results showed the high value of PROMs included in the input data even when frequently collected symptom severity reports were available. The higher importance of validated PROMs in comparison to the symptom reports was not expected from a clinical perspective. It may suggest that the overall impact of disease and treatment burden is more predictive than individual symptoms.
4. The research reported in this chapter handled the temporal irregularity of reporting by providing the model with the variable indicating days since previous report, which allowed the model to learn from the irregularity patterns. This variable was highly important for most of the analysed outcomes, according to SHAP values. This is a different approach than in similar studies [208, 265], which used synthetic data to introduce regularity. Such method introduces the risk of introducing bias in results [264].
5. The involvement of patient representatives and a clinician in the study design provided perspectives of important stakeholders and offered valuable interpretation of the results. These are necessary for any AI system to be included in clinical practice. The selection of quality of life outcomes was also previously discussed with patient representatives (Chapter 5), who mentioned that the expectation of HRQoL aspects changes throughout the treatment could contribute to treatment decision making.

Limitations

The interpretation of the research reported here should also consider the following limitations:

1. The sample size in this study is relatively small (3260 reports collected from 256 patients), which could have negatively affected the performance of LSTM, as DL models require more data than classical ML models [230] to perform well. However, this study suggests the usefulness of patient-reported data, which encourages the collection of such data on a larger scale, as an important part of clinical practice. This would enable an exploration of DL models applied to patient-reported data with bigger sample size.
2. Even though LSTM and ML models were implemented through the same pipeline, the differences included the combination of static and time-series features and handling unequal sequences of reports per patient. The discrepancies could have affected the performance of the models, but were unavoidable due to different model architectures. The number of iterations was also smaller for LSTM than ML models, which could have prevented LSTM from learning effectively. However, the trade-off between model efficiency and the robustness of the design is an unavoidable part of ML and DL comparisons.
3. The stakeholder engagement performed throughout the design of this research was highly insightful, but lacked structure. The interviews were conducted with only three patients and no thematic analyses were performed. Therefore, understanding of patient views was not derived through qualitative data analysis, but was dependant of subjective interpretation of the discussions. Nevertheless, it was an important part of the design of this research, which significantly improved model development, analysis and interpretation.

7.4.3 Conclusions

The research reported in this chapter used time-series patient-reported data (standardised PROMs and symptom severity reports), as well as static demographic and clinical information to predict HRQoL changes and hospital utilisation during chemotherapy, happening within 14 days from a symptom report, and at any time following the symptom report. LSTM model, by detecting sudden changes in data, provided better performance of admissions and triage events happening within 14 days from a symptom severity report, outperforming simpler ML models. However, admissions and triage predictions at any time during chemotherapy, and HRQoL deterioration predictions resulted with higher performance of ML models. It suggests that simpler models might be better for less complex predictions, while LSTM can detect more dynamic changes in data.

This chapter contributed to answering two of the research questions explored in this thesis:

1. **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** This chapter explored the potential of applying LSTM model to clinical, demographic, and longitudinal patient-reported data. The results suggest that often simpler ML models are sufficient for such research. However, the short-time frame of outcome required the dynamic changes detection in data, which was processed slightly better by LSTM.

2. **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Even through the clinician involvement was already present, and the outcome variables were chosen by patients, this chapter added the component of consulting the research design with patient representatives. It helped shaping this chapter to best represent patients' data for the model to learn from hidden dependencies.

This chapter presented the final primary research performed in this thesis. It helped shape the answers to research questions and used the recommendations which were built throughout the research reported in previous chapters. The following final chapter of this thesis concludes the work, summarises the findings, and discusses the overall contributions made by this thesis.

Chapter 8

Discussion

8.1 Introduction

This thesis has presented research which focuses on utilising PROMs in AI research to predict patient outcomes during chemotherapy. It has explored the methodological approaches of incorporating PROMs in AI models through utilising data collected during the eRAPID clinical trial [70]. The aims of this thesis were to propose a methodological framework by addressing four research questions:

1. **Added value of PROMs:** What is the added value of using PROMs in AI models predicting patient outcomes?
2. **Design robustness:** How do data pre-processing methods affect the results of AI models predicting patient outcomes?
3. **Stakeholder engagement:** How can patients and clinicians be involved in the design of AI models predicting patient outcomes?
4. **Data and model type:** What are the most effective AI models for predicting patient outcomes using time-series PROMs?

This chapter provides a summary of the thesis, including the main findings and key outcomes from the previous chapters. It aims to discuss contributions and recommendations drawn from the research presented in this thesis. It also provides reflections on chosen methodological approaches. Finally, this chapter outlines limitations and potential future work, which could further deepen the understanding on the topic of utilising AI and PROMs to predict patient outcomes.

8.2 Synopsis

8.2.1 Summary of Chapters

Chapter 1 introduced AI and explained the focus on this thesis to explore the type of AI models which utilise data to predict future outcomes. It provided the overview of common applications of AI in medicine, discussing challenges and opportunities. Further, the concept of PROMs was introduced, highlighting the importance of using such measures in clinical practice, which can positively contribute to shared treatment decision-making and self-management of health

conditions. It also highlighted that the increasing collection of PROMs across clinical trials and routine care provides the opportunity for this data to be used in AI models to predict patient outcomes. However, relatively small PROMs datasets, irregular reports and missing data introduce challenges for effective use of PROMs in AI predictive modelling. This chapter identified the need for methodological guidelines on utilising PROMs in AI models predicting patient outcomes and outlined the aims of the thesis.

Chapter 2 presented the scoping review, which aimed to explore the methodological choices of research using PROMs and AI to predict patient outcomes. The review involved the analysis of 94 papers selected from a literature search of six different databases. It showed that PROMs are increasingly used in AI research and have the potential to improve patient outcomes. However, the review identified four main research gaps:

1. Unclear and inconsistent added value of using PROMs across the studies.
2. Lack of robust approach for data pre-processing.
3. Limited evidence on the types of AI models which can process time-series PROMs to predict patient outcomes.
4. Rare involvement of stakeholders (patients and clinicians) in the study design.

Understanding these gaps helped establish the research questions which shaped the direction of the research reported in this thesis.

This scoping review was published in the Health and Quality of Life Outcomes journal [1] and is attached in the appendix 8.1.

Chapter 3 provided the overall research methodology. It stated research questions and explained how addressing them can improve the quality and efficacy of AI research utilising PROMs. The chapter also explained the clinical context of the research reported in this thesis. It described the eRAPID clinical trial, which was the source of the dataset and defined hospital utilisation and HRQoL, which were the outcomes predicted by AI models in this thesis. The chapter highlighted the fact that the dataset was used as a case study to draw methodological conclusions, which can answer research questions. Further, the chapter gave a brief overview of all research phases and justified methodological choices made to design the research.

Chapter 4 focused on utilising PROMs collected at a single time-point before the start of chemotherapy (baseline) to predict three hospital utilisation outcomes: hospital admissions, triage events and changes to chemotherapy. This chapter began addressing three of the research questions:

- **What is the added value of using PROMs in AI models predicting patient outcomes?** ML models were developed on six different variable sets:
 1. All available data (clinical, demographic and PROMs)
 2. Clinical and demographic data only (excluding PROMs)

3. Clinical and demographic data and PROMs collected with EQ-5D questionnaire
4. Clinical and demographic data and PROMs collected with EORTC QLQ-C30 questionnaire
5. Clinical and demographic data and PROMs collected with FACT-G questionnaire
6. PROMs only (excluding clinical and demographic data)

The comparison of the performance of models developed in these scenarios enabled understanding of the predictive value of PROMs data in general, as well as individual questionnaires. The results indicated that ML models performed better in scenarios where PROMs were available. Models trained on data that excluded PROMs were generally performing worse than models that included data from at least one questionnaire. The variable set which included EORTC QLQ-C30 questionnaire, clinical and demographic data led to better overall performance than variable sets including other questionnaires. The variable importance analyses showed that clinical and demographic data greatly contributed to the predictions, alongside PROMs variables. This suggests that PROMs can add predictive value to the models, but the inclusion of other available data is still crucial.

- **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** This chapter also explored if the techniques for addressing class imbalance in data can affect the results of ML models. Therefore, three different techniques were used and compared:

1. **Original dataset:** Fitting models to original data, without balancing train and test sets.
2. **Balanced dataset:** Balancing the dataset prior to splitting data to train and test sets.
3. **Partially balanced dataset:** Balancing train set, leaving test set unbalanced (original).

Each of these methods was used in the similar studies predicting patient outcomes from PROMs, as stated in the scoping review (Chapter 2). Since the use of these techniques is inconsistent in literature, the clarity of how they impact the results is needed. The comparison of the performance of models developed and evaluated on data affected by these methods showed a large discrepancy in the results. Balancing all data led to a higher model performance than balancing only train set and leaving all data imbalanced. Therefore, this method may introduce bias in model evaluation and provide unrealistic predictions, as the real-world data is rarely perfectly balanced. Performance of models developed with balanced train set only provided slightly better performance than models developed on original data in predicting admissions. There was no noticeable difference for other outcomes. However, the model calibration was improved for partially balanced dataset. This suggested that balancing train set can support models to learn more effectively from the data and prevent the bias towards any particular class.

- **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Despite the lack of patient involvement at this stage of the thesis, patients' views were represented through the engagement of the PCOR group and a clinical oncologist throughout the design of research reported in this chapter. The collaboration with these stakeholders helped improve the following aspects of the research design:
 1. The stakeholders considered variable importance analysis very important to understand the decisions made by the models and assess their explainability. Therefore, the variable importance was added to the analysis in this chapter.
 2. The stakeholders expressed concerns regarding evaluating model performance on data affected by balancing techniques. They suggested that the model evaluation would be the most reliable when test set is not affected by any pre-processing methods. However, they did not mind what types of data the models are trained on. Therefore, the performance of models developed on partially balanced dataset was compared with balanced and imbalanced dataset.

This part of the research was published in the JCO Clinical Cancer Informatics journal [2]. The paper is attached in the appendix 8.2.

Chapter 5 also utilised PROMs data collected at one time-point (baseline of chemotherapy). However, the outcomes in this chapter were HRQoL changes. These outcomes were selected following a consultation with patient representatives. Four ML models predicted changes to self-reported mobility (EQ-5D), physical well-being (FACT-G), social functioning (EORTC QLQ-C30), role functioning (EORTC QLQ-C30), and usual activities (EQ-5D) from baseline to 6, 12 and 18 weeks throughout chemotherapy. The research questions addressed by this chapter were:

- **What is the added value of using PROMs in AI models predicting patient outcomes?** Despite utilising only baseline PROMs as predictor variables, this chapter explored how longitudinal PROMs can be utilised to predict HRQoL changes. Due to the time-series properties of the dataset, it was possible to calculate the changes from the baseline to three different time-points throughout the chemotherapy. Generally, predictions for 18-week and 12-week change had a better performance than 6-week change.
- **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** Alongside the ongoing involvement of a clinical oncologist, patient representatives were consulted during the study design. The most important input from the stakeholders was selecting the variables, which they consider important. HRQoL changes during chemotherapy, especially the ones exploring social context and daily activities, are less expected and rarely discussed in clinical setting, compared to the risk of hospitalisations or pain. Therefore, patients mentioned the need to understand how their life would change. The inclusion of outcomes important and relevant to patients improved the feasibility of this research and helped ensure that the actual need is addressed.

The research conducted during this chapter was published in the proceedings of the 22nd International Conference on Artificial Intelligence in Medicine in Salt Lake City, USA, July 2024 [3]. The paper is attached in the appendix 8.3.

Chapter 6 was the first chapter of this thesis to explore predictions of patient outcomes based on longitudinally-collected PROMs. The symptom severity reports collected weekly throughout the chemotherapy were also utilised for the first time. This chapter predicted hospital utilisation and HRQoL outcomes, as both were considered important by stakeholders. The hospital utilisation outcomes were split into short-term (happening within 14 days from a symptom severity report) and long-term (happening at any time during chemotherapy) predictions. The long-term predictions are standard types of prediction in similar studies and can inform the treatment decisions and help prepare for potential deterioration. The short-term predictions could support timely prevention of adverse events, and are rarely explored in literature. As in previous chapters, ML models were developed. Despite the popularity of using DL models for processing time-series data, ML models were the most common method to predict patient outcomes from longitudinal PROMs. Therefore, this chapter explored the ML models ability to predict both short- and long-term predictions and understand opportunities and limitations of using these methods. Chapter 6 addressed all of the research questions:

- **What is the added value of using PROMs in AI models predicting patient outcomes?** Including time-series patient-reported data and predicting the same outcomes in the previous chapters enabled the comparison of the results and understand the value of longitudinal PROMs and symptom severity reports. Furthermore, the value of standardised PROMs was explored through the comparison of the performance of models including and excluding them, when symptom severity reports were available. The results suggested that including time-series patient-reported data leads to better performance of all outcomes. Furthermore, adding PROMs to symptom severity reports also improved the model performance.
- **How do data pre-processing methods affect the results of AI models predicting patient outcomes?** A very important aspect of this thesis is protecting a test set from being affected by pre-processing methods. This is due to the importance of evaluating models on data which is as similar as possible to real-world data, which could increase the relevance of this research to clinical setting. Therefore, this chapter used a nested CV pipeline for model development. This pipeline is considered an improvement from the previous chapters, which used CV only for hyperparameter selection. The nested CV provides a more robust way of model tuning and minimises bias arising from data leakage, as all data pre-processing is completed within both inner and outer CV loops [242, 246].
- **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** The involvement of patient representatives and a clinical oncologist continued throughout the the design of research reported in Chapter 6. Due to the large volume of variables, it was

very helpful to discuss their relevance with a clinical oncologist. Selecting only clinically important variables helped prevent noise in data, and ensured that models learn from relevant information. Furthermore, the time window of 14 days for short-term prediction was selected after consulting a clinical oncologist. The individual interviews with patient representatives further improved the study design. The discussions enabled understanding of potential reasons for irregularity in data. This helped decide to include variable which informs the model of the time between two reports and as a result learn from the irregularity.

- **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** This chapter explored four different ML models applied on time-series patient-reported data to predict outcomes. The overall performance of long-term prediction was better than average performance reported in studies analysed in the scoping review (Chapter 2). The recall and sensitivity of the models were high, suggesting that both classes were accurately predicted. Therefore, this chapter suggests that ML methods can be sufficient for predicting outcomes which are not limited within any time frame. However, the ML models were not successful in short-term predictions. The limited time for the outcome to happen (14 days) provided an additional challenge to the prediction task, which suggested that more complex models might be better suited to such outcomes.

A part of the research reported in this chapter was published in a proceedings of a Medical Informatics Europe (MIE) Conference in Glasgow, Scotland, UK, May 2025 [4]. The paper is attached in the appendix 8.4.

Chapter 7 was the final chapter reporting the original research conducted for this thesis. It explored the performance of LSTM model (a type of DL models) to predict the same outcomes as in Chapter 6. It addressed two of the research questions:

- **What are the most effective AI models for predicting patient outcomes using time-series PROMs?** Comparison of the performance of LSTM model and ML models developed in Chapter 6 enabled understanding which methods are the most effective for short- and long-term predictions when time-series patient-reported data are used. ML models outperformed the LSTM model for long-term predictions. However, LSTM slightly improved the predictions of hospital admission happening within 14 days from the symptom report. This suggests that more complex DL models might provide a better solution for timely processing of temporal dependencies. However, for simpler predictions, using ML models might be a sufficient method for processing time-series PROMs and symptom severity reports.
- **How can patients and clinicians be involved in the design of AI models predicting patient outcomes?** The discussion with patient representatives regarding the irregularity of reported symptoms was extremely valuable for this chapter. Due to the LSTM models requiring equal sequence of data for each patient, handling irregular data was challenging. Patient representatives suggested that the reasons for missing reports or too frequent report might be

related to the well-being and a health state on a particular day. Therefore, the common approach of aggregating and imputing data would remove the important information from the dataset. Following the consultation, padding technique with masking was implemented, which created the equal sequences without altering the data. A variable which informed of the time between reports was kept to further facilitate learning from the temporal irregularity.

A part of the research reported in this chapter was published in the proceedings of the 13th IEEE International Conference on Healthcare Informatics in Pavia, Italy in June 2025 [5]. The paper is attached in the appendix 8.5.

8.2.2 Key Findings

The key findings of the research reported in this thesis have helped address the gaps which were identified through a scoping review presented in Chapter 2:

1. Added value of PROMs:

- Including PROMs in AI models has the potential to improve model performance for predicting patient outcomes.
- Clinical and demographic variables were important for predictions of patient outcomes. Models trained on PROMs as stand-alone input data did not outperform models trained on all available variables.
- Using PROMs collected longitudinally throughout chemotherapy improved the performance of models predicting patient outcomes, compared to using baseline PROMs.
- Longitudinal standardised PROMs questionnaires improved performance of models even when symptom severity reports were available.

2. Design robustness:

- Techniques for handling class imbalance in data can alter the results.
- Using balancing methods prior to splitting data to train and test sets can improve the model performance. However, this can be a result of data leakage and providing a model with unrealistic data for evaluation of performance.
- Using nested CV is a preferred pipeline for model development and evaluation, as it protects test set from data leakage and data pre-processing, which is important for consulted stakeholders.

3. Stakeholder engagement:

Figure 8.1 summarises what aspects of study design can be improved by the involvement of patient representatives and clinicians. Stakeholder engagement supported:

- Selection of relevant input variables.
- Methodological choices of data pre-processing techniques which do not alter important and relevant patterns in data.

- Understanding the need of using explainability techniques to help understand model predictions.
- Robustness of models through understanding the importance of protecting test set from data leakage and data pre-processing methods.
- Ensuring that predicted outcomes are relevant and address real clinical need.

4. Data and model type:

- ML models trained on time-series PROMs and symptom severity reports provided a better performance than ML models trained on PROMs collected at one time-point.
- ML models outperformed LSTM models predicting hospital utilisation and HRQoL changes at any time throughout chemotherapy.
- LSTM model outperformed ML models predicting hospital utilisation events happening within 14 days from a completed symptom severity report.

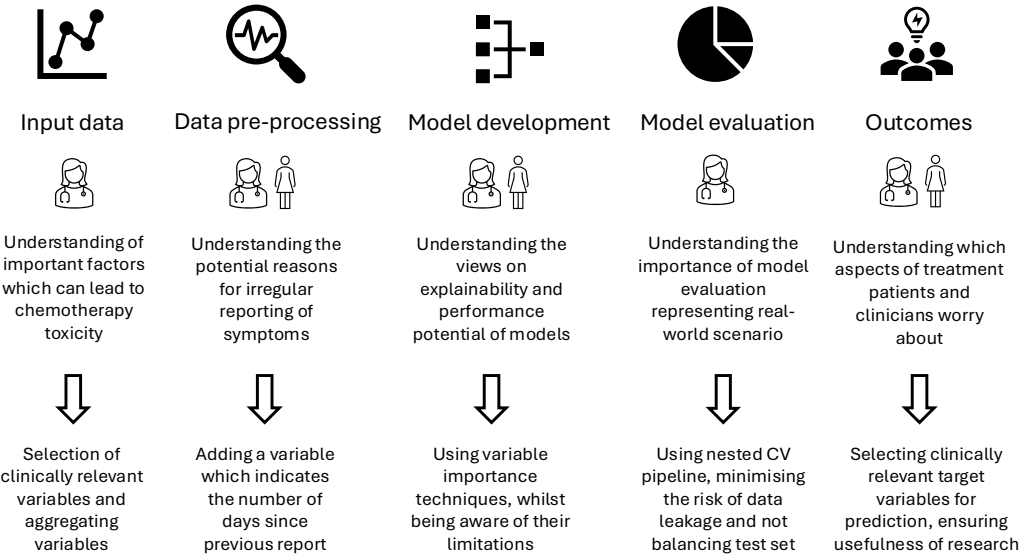


Figure 8.1: Aspects discussed with stakeholders throughout the study design, and the impact stakeholder engagement had on each part of methodology.

8.3 Recommended Framework

There is a growing interest in using AI in medical research [31], but the developed AI systems are rarely applied to clinical practice. Poor reporting and quality of research can lead to inadequate evaluation of models and prevent reproducibility of studies [266]. It becomes increasingly important to follow evidence-driven

methodological guidance, which can support maintaining high quality of research. For instance, the TRIPOD+AI checklist [203] (used in this thesis) helps ensure appropriate and consistent reporting of medical AI research. International Journal of Medical Informatics (IJMEDI) checklist is another useful self-assessment form which was created to increase the quality of medical AI research [266]. These are very helpful tools, which can improve transparency, reproducibility and standard of AI methods used for predictions in healthcare. However, these checklists are generic and do not differentiate between data types used as input variables.

PROMs datasets introduce challenges specific to data reported by patients, with missing PROMs data being a common and important issue, requiring additional attention [73]. There is existing guidance on using PROs in clinical trials. For example, CONSORT-PRO supports an adequate reporting and justification of the instruments used, analytical techniques, and outcome measures for clinical trials utilising PROs [50, 51]. Exploration of missing data handling is also covered by CONSORT-PRO and SISAQOL guidelines [51, 267], with the consideration that reports missing not in random may be related to the deterioration or improvement in health status. However, in studies reporting AI research utilising PROMs, the challenges of irregular reporting (i.e. missing or too frequent reports) are handled inconsistently, as shown in Chapter 2. The lack of guidelines on how to handle PROMs data and poor adherence to existing checklists revealed that there is a need to provide methodological recommendations for using PROMs and AI to predict patient outcomes.

The main goal of this thesis is to introduce a framework for developing and evaluating clinically relevant AI models applied on PROMs. The framework has been derived from the research reported in Chapters 4, 5, 6, 7, addressing the gaps identified in the scoping review (Chapter 2). The framework is presented in Table 8.1 and provides recommendations for utilising PROMs and AI models to predict patient outcomes. It considers the type of PROMs questionnaires, class imbalance handling, irregularity of reporting, model development, evaluation, and the involvement of relevant stakeholders. The framework should be considered an addition to the available checklists for medical AI research.

8.4 Original Contributions

As described in Chapter 3, the research reported in this thesis should be considered as case study, due to using only a single dataset and lack of external validation of the results. The contributions which relate to the methodology which was assessed in this thesis can be considered as generic contributions. They can be applicable for any research using PROMs and AI methods to predict patient outcomes. However, this thesis resulted in further contributions which are specific to the predicted outcomes of hospital utilisation and HRQoL changes during chemotherapy.

8.4.1 Generic Contributions

This section outlines the contributions of the work reported in this thesis, which can be applicable to wider research predicting patient outcomes using PROMs and

Table 8.1: The framework for developing and evaluating clinically relevant AI models applied on PROMs.

Recommendation	Justification
If available, use standardised PROMs questionnaires.	As reported in scoping review (Chapter 2), inconsistent selection of PROMs introduces a challenge for evaluation of patient-centric AI models in healthcare. The research in this thesis also suggests that standardised PROMs can increase model performance, even when symptom severity reports are available.
If available, use PROMs collected longitudinally.	Scoping review (Chapter 2) has shown that frequently availability of time-series PROMs were not utilised and only reports from one time-point were considered for analysis. However, the research reported in this thesis showed that time-series PROMs can improve the performance of predictions of the same outcomes.
Use clinical and demographic variables alongside PROMs	The variable importance analysis and comparison on models built on different variable sets showed that clinical and demographic variables are important for predictions. PROMs provide a valuable addition, increasing model performance.
If there is a class imbalance in data, do not balance test set	To ensure robust model evaluation, test set should not be altered by any data pre-processing techniques. The research reported in this thesis suggests that balancing data prior to train/test split can lead to bias.
Perform all data pre-processing steps within nested cross-validation loops	Nested cross-validation pipeline is a robust solution for developing and evaluating models, preventing data leakage and protecting test set from data pre-processing methods. It was important to consult stakeholders that the test set is as similar as possible to real-world data.
If using time-series PROMs, preserve irregularity information in data.	Time-series data is frequently heavily pre-processed to alter the sequence of time-steps. It often involves aggregating reports which are completed too often and synthetically imputing missing reports. However, the consultation with stakeholders indicated that irregular pattern in PROMs data can hold information about patients' health status. Therefore, protecting the irregularity in data (for example through adding a variable indicating time between reports) can provide models with better opportunities to learn.
Consult stakeholders regarding all stages of study design.	The design of research reported in this thesis has been improved due to the consultations with stakeholders. It helped understand the properties of patient-reported data, the relevance of outcomes, the views on explainability of the model. Engaging patients and clinicians has the potential to improve the public trust in medical AI research. If possible, consult diverse group to increase the representativeness of population views.
When processing time-series PROMs, use ML models as a default.	Due to inherently small and non-complex PROMs datasets, ML models have the potential to perform well predicting outcomes from time-series PROMs. This thesis has shown that ML models outperformed LSTM in most scenarios. This is also common in other research using tabular data. ML models are simpler than DL models, are more efficient, faster, explainable and require less computational power.
Use DL approaches only when ML is not performing sufficiently well.	There is a growing trend of using DL methods for prediction tasks due to their potential to detect anomalies in complex, high-dimensional data. However, the interpretability and explainability of those models are compromised. They also require high energy consumption and computational power. This thesis has shown that LSTM only had a small advantage over ML models in short-term predictions. Therefore, using DL approaches cautiously and only in situations where ML fails, can prevent unnecessary costs and waste of resources. It can also support the need for transparency, explainability and interpretability of AI models.

AI.

The first contribution is the critical assessment of the added value of PROMs in predicting patient outcomes. After thorough search of literature (Chapter 2), it can be concluded that this is the first work which critically assesses the utility of PROMs in AI predictive models. Not only has this thesis provided comparisons of models which included and excluded PROMs from the analysis, but also looked into specific types of PROMs. Firstly, the predictive value of three different questionnaires (EQ-5D, EORTC QLQ-C30, FACT-G) was assessed, which helped understand how they impact model performance individually. Secondly, the outcomes were predicted by baseline PROMs collected at one time-point, as well as by time-series PROMs and symptom severity reports. The predictive value of the two types of longitudinal PROMs was also assessed. The comprehensive and critical analysis of PROMs used in AI predictive models provides an insight into utilising and collecting PROMs. It encourages using standardised PROMs questionnaires if available, as the symptom severity reports did not add much predictive value, contrary to clinical expectations. It also suggests that PROMs are only effective when used alongside clinical and demographic variables.

The approach for data pre-processing of PROMs for AI research provides another contribution of this thesis. Class imbalance in data is a frequent challenge faced by research using classification models. The scoping review (Chapter 2) revealed inconsistency in methods handling class imbalance. One of the frequently used methods was balancing dataset before splitting into train and test sets. The comparison of this technique with a technique that leaves test set original showed that the choice of balancing methods can impact the results. This suggests that there is a risk of bias and data leakage when test set is altered. This thesis contributed to understanding this risk also through discussing it with patients and clinicians. The stakeholders believed that the models should be evaluated on data as similar as possible to real-world data. Therefore, the suggestion to use nested CV design contributes the approach which can increase public trust in AI models used in medicine and provides the opportunity to develop models which are clinically relevant.

The insights from stakeholder engagement is another contribution of this thesis. The scoping review (Chapter 2) showed that only 3 out of 94 papers reported clinicians' inputs in the study design. No papers reported the involvement of patients. This suggests that stakeholders' views are rarely taken into consideration when using patients' data for predicting outcomes. Even though there are several limitations of the approach that this thesis takes to involve patients and clinicians, this thesis has shown how stakeholders can be involved in each stage of research design (Figure 8.1), based on described examples. It also provides insights on how this collaboration can improve the study design through understanding the properties of data reported by patients, the relevance of outcomes, the views on explainability, and the importance of evaluating model on adequate data. Through these recommendations, this thesis contributes to improving public trust in medical AI and ensuring clinical relevance of research. Other best-practice recommendations should also be followed to improve the public involvement and engagement in health research [268].

Another contribution of this thesis is a critical assessments of several AI methods predicting patient outcomes from time-series PROMs. Exploring various ML models and DL models applied on time-series PROMs for both long-term and short-term prediction provides the comprehensive understanding of complexity of a task which might require using DL models. This thesis has shown that ML models were sufficient to predict most outcomes from time-series data and the advantage of LSTM in short-term predictions was not significant. Therefore, contrary to current trends to use computationally expensive DL models [269], this work suggests using ML models as a default for tasks using time-series PROMs with limited complexity.

8.4.2 Contributions Related to the Specific Clinical Context for this Research

Due to using a single dataset, this thesis cannot draw conclusions on predicted outcomes, as the models are not externally validated. However, this thesis has provided some contributions, which can impact the direction of similar work.

This thesis provides insights into using PROMs collected through clinical trial provide. According to a systematic review on trials of artificial intelligence health technologies [74], PROMs collected through clinical trials are rarely utilised as inputs or outputs in AI models. Out of trials that used at least one PROM, visual analogue scale, patient-reported experience measure, or usability measure, only 4% were utilised as inputs, and 0.3% as both inputs and outputs. Considering such small number of PROMs collected through clinical trials being utilised in AI predictive models, this thesis provides examples on how such data can be applied in similar research. It also encourages the use of PROMs which are increasingly being collected in routine care as a result of mobile technology and integration in EHR [74]. It is a timely contribution, considering the increasing use of digitally collected PROMs to support and improve individual patient care. Due to mobile and online technology, the collection of PROMs is expected to increase dramatically in the future with the wider use of digital health care. This thesis uses data collected in a clinical trial of a digital intervention mimicking routine care with added symptom reports and the results provide an example of how such data can be utilised to predict health care and patient outcomes.

Predicting HRQoL changes during chemotherapy can be considered a contribution to the field of oncology. There is existing evidence of ML predicting chemotherapy toxicity, but studies often focus on acute hospital utilisation during treatment, rather than HRQoL [252]. Research foreseeing HRQoL during chemotherapy tends to use statistical modelling, rather than explore ML predictive opportunities [236]. Furthermore, the outcomes are usually limited to a specific time during the chemotherapy, without investigating the changes from early to late stages of the treatment. Understanding factors related with HRQoL fluctuations during chemotherapy could personalise cancer care through enabling informed shared decision-making and early preparation for potential adverse effects. This thesis also highlights the need of exploring HRQoL changes through conversations with patient representatives. The work reported in this thesis considers HRQoL

outcomes, which contributes to patient-centred research by providing examples of how PROMs can be utilised as outcomes.

Short-term predictions of hospital utilisation provide another contribution. Even though the results of short-term hospital utilisation predictions should be validated by further research, this thesis highlights the importance of such outcomes. A successful system that can alert to potential complications during chemotherapy in a rapid manner, prompting immediate hospital contact can be extremely useful to prevent serious complications or hospitalisation. This thesis introduces the challenge of predicting outcomes within the short time from a symptom severity report and provides guidance on how to handle similar data, especially when combining static and temporal variables in LSTM model.

Furthermore, this thesis contributes the variable importance analysis. The exploration of how variables in the dataset contribute to predictions of outcomes by different models is another contribution. Variable importance led to the understanding that clinical and demographic data should not be omitted from analysis as they play an important role in predictions alongside patient-reported data. Furthermore, it provided interesting insights into the differences in symptom reports and validated PROMs as predictors. Contrary to clinical expectations, LSTM for short-term predictions of hospital utilisation used mainly PROMs data, rather than symptom reports. This suggests that the model is considering functional status and well-being as more predictive of hospital utilisation than symptoms of toxicity happening closer to the time of admission. It is important to note the limitation of SHAP analysis used for variable importance analysis, which is assuming independence of features [270]. Since symptom reports are related to overall health and well-being, they might have been omitted by the explainability analysis.

8.5 Strengths

This thesis has a number of key strengths. The design of research reported in this thesis was based on the thoroughly conducted scoping review. Systematically analysing the studies enabled deeper understanding of the field, minimising the risk of missing relevant papers. The research questions developed through identified gaps provide a strong justification of the need to explore methodological approaches in utilising PROMs and AI to predict patient outcomes.

Furthermore, this thesis advocates for patient-centric approach in two main ways. Firstly, it considers patients' perspective through using data reported by patients. This leads to the patients' voice being "heard" by AI models, predicting outcomes relevant to them. Secondly, the active engagement of a clinical oncologist and patient representatives throughout the study design enabled understanding what is truly important to patients.

Using checklists, such as PRISMA for the literature review and TRIPOD+AI for original research, ensures that no important information is missing and that the thesis follows the state-of-the-art approach for robust reporting.

Attempts to interpret and explain ML and LSTM models should be considered as another strength of this thesis. Understanding the decision making process of models is inevitable part of increasing trust in medical AI research.

8.6 Limitations and Future Work

There are several limitations which point at future research. The main limitation of this thesis is the use of single dataset collected as a part of a digital health intervention trial. To mitigate this limitation, the thesis has followed a rigorous approach and has ensured that the key decisions are grounded on a state-of-the-art approach to allow generic contributions (see Section 8.4.1). However, due to inclusion and exclusion criteria of the clinical trial, the sample is not representative of the population. This is because people who are willing to participate in research often have characteristics which differ from an average person. For example, the majority of participants of clinical trials are white and male [271]. Education, employment status, insurance coverage, and English proficiency are also predictors of participation in clinical trial [272]. Furthermore, participants tend to have different psychological characteristics and are more curious and self-efficacious, less anxious and reported a higher level of social support [273]. To increase the potential of such models being applied in clinical practice, validating the model on external datasets should be considered in the future. Such datasets should ideally be collected in a similar context, multi-centre national or international settings and provide information of all demographic characteristics. The external validation study should also consider comparing the predictive performance in different subgroups to test for bias. This is an important ethical consideration that can prevent models from replicating existing inequalities in healthcare. External validation was not performed in this thesis due to the methodological focus on applying AI models on PROMs and lack of access to similar datasets. It is also helpful to ensure following well-established guidelines for tackling bias in health datasets, such as recommendations from STANDING Together initiative. [274].

The comparison of model performances in this thesis was mainly done by adding and taking away variables, which is another limitation of this thesis. This is a common approach widely used in similar studies [6] and it allows direct comparison of model performance. However, AI models tend to be more accurate when trained on more data [275]. In this thesis, the models which were trained on more variables often led to better performance than models trained on fewer variables. For example, PROMs and clinical data led to better performance than just clinical data. Therefore, there is a risk that models including PROMs perform better due to a larger dataset provided for training, rather than due to the predictive value of PROMs. This limitation was mitigated through looking at the performance holistically, including variable importance analyses, which assess the impact of each variable separately. Understanding how each variable contributed to predictions provides additional insight into the predictive value of PROMs and the understanding of which questionnaires have a potential to improve model performance. Another way this limitation could have been mitigated would be to select the same number of variables from each compared dataset. However, with

the focus on clinical relevance, the approach in this thesis was to discuss variable selection with a clinical oncologist, rather than follow solely data-driven process.

The thesis embarked upon the challenge to engage stakeholders (patients and clinicians) in the design of AI models. However, despite stakeholder engagement activities providing a lot of insights into the design of the research, the involvement of patients and clinicians could be improved. To draw reliable conclusions on stakeholders' views on the design, more structured activities, such as scripted focus groups, recorded sessions, thematic analysis should be performed. However, such research would require a separate ethics application and a shift of focus from the methodological contributions. Therefore, based on the availability of research advisory groups and time, in total six patient representatives provided consultations for the research reported in this thesis (three for Chapter 5 and three for Chapter 7). Furthermore, these patients were actively taking part in consultations for research. This suggests that they might not have been representative of the population, having more knowledge on clinical research than an average patient and overall positive views on AI in healthcare. A larger and more diverse sample of the patients would have provided better understanding of general views on utilising AI in medicine. This limitation also relates to clinician involvement. The PCOR group and the clinical oncologist, who was a project supervisor are researchers who have overall positive views on PROMs and AI. Future studies should consider conducting qualitative research on stakeholder engagement throughout study design for utilising AI and PROMs and involve patients and clinicians who represent diverse perspectives on such research.

8.7 Reflections

The research described in this thesis has been conducted for over three years. During this time, I have had an opportunity to develop as a researcher and make methodological choices which reflect the best of my knowledge and experience at a particular time. Inevitably, the methods have been gradually changing throughout this thesis. For instance, the choices of models and evaluation metrics in Chapters 5 and 6 differed from Chapter 4. This is because Chapter 4 was motivated by preliminary results of the scoping review. After thoroughly quantifying all of the most common approaches in research, the model and evaluation metric selections changed. Furthermore, the choice of a nested cross-validation pipeline for model development in Chapter 6 was made following a reviewer's comments of the manuscript submitted to the MIE conference [4]. This was a valuable lesson, which prompted more literature searches on the benefits of this method and its utilisation in the research.

The factor that has influenced my development the most as a researcher was stakeholder engagement. Listening to people who have undergone chemotherapy shifted my focus from trying to maximise the technological novelty of this research to ensuring feasible design which can address actual clinical needs. Due to the availability of only one dataset, it was clear that the models reported in this thesis will not be applied in clinical practice. However, through the patient engagement, this thesis could provide the guidance on how to strengthen the methods, minimis-

ing bias, and hopefully increasing a clinical relevance of research.

In the recent years, there has been a rapid growth of research proposing various opportunities for applying AI in different health domains. On the other hand, there is an increasing awareness of ethical considerations regarding the risks of bias deepening health inequalities and the negative impact of AI on the environment [276]. I hope that the narrative presented in this thesis showed the enthusiasm about the great potential of AI's role in patient-centred research, whilst providing cautious and critical assessment of the explored methods. The awareness of bias and attempts to mitigate it have been an integral part of this thesis. However, the lack of ethnicity information in the data prevented a thorough exploration of bias. This does not form any immediate risk, as the models reported in this thesis are not intended to be implemented in clinical practice. Furthermore, the considerations of computational power of DL models and their impact on the environment have influenced the recommendations on handling time-series PROMs. This thesis suggests to only use DL models if ML does not achieve a satisfactory performance. It is not the aim of this thesis to discourage from developing and using new technological advancements, but to prompt the reflection when selecting model types, with the consideration of the environment. If a task can be solved with a simpler method, there is no need for a computationally expensive one just because it is more novel.

The growing interest in collection and use of PROMs in clinical practice, clinical trials, and medical AI research suggest that the views of patients are increasingly listened to. AI models which take these views into consideration have the potential to provide more accurate outcomes and, more importantly, show to patients that their voice matters and that they are the experts in their own health status. This thesis has provided recommendations on how to best utilise patients' voice collected through PROMs. Hopefully, in the future the use of patients' voice through including PROMs in AI predictive models and wider stakeholder engagement will be a common approach. With the technological advancements, routine PROMs collection can provide opportunities for utilising such data in wide range of medical AI research in various healthcare domains and can be combined with other data, e.g. wearable sensors.

8.8 The Future of AI and PROMS

The field of AI has increasingly been applied in medical research, revolutionising methods for early detection of illnesses, personalising treatments and reducing hospital burden. The growing interest in collection of PROMs provide high-value datasets, which have a potential to be utilised in AI models predicting patient outcomes. The work reported in this thesis has summarised how AI and PROMs have been used, as well as suggested a framework for consistent and responsible application of these tools.

The focus on AI integration in medicine is relevant to current trends in digital transformations of healthcare systems around the World. For example, in the UK, the National National Health Service (NHS) 10-year plan introduces three big

shifts: reducing the reliance on hospitals, providing better technological solutions, and emphasising prevention over treatment [277]. The insights from the work reported in this thesis directly align with the vision for the future of the NHS.

To begin with, the focus on PROMs itself can help reducing burden on hospitals. Evidence suggests that real-time monitoring with electronic collection of PROMs can improve well-being and self-efficacy without increasing hospital workload [70]. Applying AI on data collected via PROMs can further help reducing over-reliance on hospitals. AI systems have a potential to flag early warning patterns in longitudinal symptom reports, as explored in Chapters 6 and 7 of this thesis. AI could be utilised to suggest preventative measures and provide self-management advice to reduce risk of emergency visits and hospitalisations.

This thesis has also shown how new technological solutions can be implemented to support the management of health services. It has explored both traditional ML models and more complex DL model to predict patient outcomes during chemotherapy. However, the main focus has been on improving the process of AI modelling to become less biased, more responsible, and relevant to problems experienced by people with lived experience of cancer. The current limitation is that medical AI often lacks the direct link to patient experience [72]. Whilst AI is becoming increasingly multimodal, autonomous and generative, PROMs can offer personalised decision support, which accounts for quality of life, well-being and subjective reports of health status. This can improve outcomes of the AI, making the technology fit for purpose.

Emphasising prevention over treatment is another important shift described in the NHS 10-year plan. Forecasting patient deterioration and risk of hospital utilisation can support prevention of these outcomes. The research reported in this thesis has shown that AI models are able to predict these outcomes, and that PROMs can further improve these predictions. PROMs can hold information on patterns of fluctuating symptoms and quality of life aspects, and as a result enable prevention of deterioration in health status.

Aligning the outcomes explored in this thesis with the NHS 10-year plan has shown that the future of medical research is likely to include AI, and that PROMs can help improve the utilised models. However, using PROMs data does not automatically mean that the models consider all patients' characteristics. The challenge of ethical and inclusive medical AI still stands. When allowing patients' perspectives to be used in AI models, it is important to remember that this perspective is only collected from people who are willing to provide it. Patients who decide to take part in clinical trials rarely represent the entire population. Usually, these patients are educated, positive about research and often have characteristics which differ from an average person [271, 272]. Therefore, alongside the guidance from this thesis, it is important to follow recommendations of initiatives, which tackle bias and promote transparency in health datasets, such as STANDING together initiative [36].

There is also no ethical AI without involving patients in the process of cre-

ating it. The literature review presented in Chapter 2 exposes a lack of stakeholder engagement in previous work. This thesis has provided only few examples of patient and clinician involvement in AI research, so it should not be understood as a gold standard. However, it has suggested the direction of stakeholder engagement, which future research should follow to increase relevance of AI in medicine.

The future of utilising AI and PROMs in medical research is very optimistic. If used responsibly, these tools have a potential to increase a success of the NHS and other health services to digitalise healthcare, reduce hospital burden and improve prevention. This work, based on thorough literature review and primary research, identifies the main issues preventing the application of AI and PROMs in medicine: unclear value of PROMs, lack of robust methodological frameworks for pre-processing of static and time-series data, and limited stakeholder engagement. Therefore, this thesis recommends measures which can mitigate these limitations and help improve adoption of such systems. Used together with other well-established guidelines, the insights from this thesis can help address the needs of health services for digital and efficient systems focused on prevention.

Bibliography

- [1] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, and K. Absolom, “Using artificial intelligence to predict patient outcomes from patient-reported outcome measures: A scoping review,” *Health and Quality of Life Outcomes*, vol. 23, no. 1, p. 37, Apr. 2025, ISSN: 1477-7525. DOI: 10.1186/s12955-025-02365-z. Accessed: Apr. 24, 2025. [Online]. Available: <https://doi.org/10.1186/s12955-025-02365-z>.
- [2] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, and K. Absolom, “Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From Patient-Reported Outcome Measures,” *JCO Clinical Cancer Informatics*, no. 8, e2300264, Apr. 2024, Publisher: Wolters Kluwer. DOI: 10.1200/CCI.23.00264. Accessed: Apr. 26, 2024. [Online]. Available: <https://ascopubs.org/doi/10.1200/CCI.23.00264>.
- [3] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, and K. Absolom, “Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes During Chemotherapy,” en, in *Artificial Intelligence in Medicine*, J. Finkelstein, R. Moskovitch, and E. Parimbelli, Eds., Cham: Springer Nature Switzerland, 2024, pp. 101–116, ISBN: 978-3-031-66538-7. DOI: 10.1007/978-3-031-66538-7_12.
- [4] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, K. Absolom, and S. D. Relton, “Machine Learning Models Predicting Hospital Admissions During Chemotherapy Utilising Longitudinal Symptom Severity Reports and Patient-Reported Outcome Measures,” en, in *Intelligent Health Systems – From Technology to Data and Knowledge*, IOS Press, 2025, pp. 178–182. DOI: 10.3233/SHTI250297. Accessed: Jul. 25, 2025. [Online]. Available: <https://ebooks.iospress.nl/doi/10.3233/SHTI250297>.
- [5] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, K. Absolom, and S. D. Relton, “Time-Series Patient-Reported Data and LSTM Predicting Hospital Utilisation During Chemotherapy,” in *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, ISSN: 2575-2634, Jun. 2025, pp. 228–234. DOI: 10.1109/ICHI64645.2025.00034. Accessed: Aug. 11, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/11081675>.
- [6] J.-a. Sim et al., “The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: Applications of machine learning,” en, *Scientific Reports*, vol. 10, no. 1, p. 10 693, Jul. 2020, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-020-67604-3. Accessed: Sep. 25, 2025. [Online]. Available: <https://www.nature.com/articles/s41598-020-67604-3>.

-
- [7] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” eng, *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019, ISSN: 1942-4787. DOI: 10.1002/widm.1312.
- [8] D. Gajjar, “Artificial Intelligence: An explainer,” en-GB, Jun. 2025. Accessed: Aug. 6, 2025. [Online]. Available: <https://post.parliament.uk/research-briefings/post-pb-0057/>.
- [9] X. Du-Harpur, F. Watt, N. Luscombe, and M. Lynch, “What is AI? Applications of artificial intelligence to dermatology,” *British Journal of Dermatology*, vol. 183, no. 3, pp. 423–430, Sep. 2020, ISSN: 0007-0963. DOI: 10.1111/bjd.18880. Accessed: Jan. 3, 2024. [Online]. Available: <https://doi.org/10.1111/bjd.18880>.
- [10] C. Janiesch, P. Zschech, and K. Heinrich, “Machine learning and deep learning,” en, *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, ISSN: 1422-8890. DOI: 10.1007/s12525-021-00475-2. Accessed: Apr. 7, 2025. [Online]. Available: <https://doi.org/10.1007/s12525-021-00475-2>.
- [11] P. P. Shinde and S. Shah, “A Review of Machine Learning and Deep Learning Applications,” in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBE)*, Aug. 2018, pp. 1–6. DOI: 10.1109/ICCUBE.2018.8697857. Accessed: Apr. 7, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8697857>.
- [12] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, “Bias and Class Imbalance in Oncologic Data—Towards Inclusive and Transferrable AI in Large Scale Oncology Data Sets,” en, *Cancers*, vol. 14, no. 12, p. 2897, Jan. 2022, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-6694. DOI: 10.3390/cancers14122897. Accessed: Jun. 8, 2023. [Online]. Available: <https://www.mdpi.com/2072-6694/14/12/2897>.
- [13] J. Bernett et al., “Guiding questions to avoid data leakage in biological machine learning applications,” en, *Nature Methods*, vol. 21, no. 8, pp. 1444–1453, Aug. 2024, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-024-02362-y. Accessed: Sep. 19, 2025. [Online]. Available: <https://www.nature.com/articles/s41592-024-02362-y>.
- [14] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. Accessed: Sep. 12, 2025. [Online]. Available: <https://dl.acm.org/doi/10.1145/2939672.2939785>.
- [15] I. B. Wilson and P. D. Cleary, “Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes,” eng, *JAMA*, vol. 273, no. 1, pp. 59–65, Jan. 1995, ISSN: 0098-7484.
- [16] E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, Eds., *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*, English. Springer Nature, 2023, Accepted: 2023-01-20T16:54:39Z. DOI: 10.1007/978-981-19-5170-1. Accessed: Apr. 10, 2025. [Online]. Available: <https://library.oapen.org/handle/20.500.12657/60840>.

-
- [17] B. H. Shekar and G. Dagneu, "Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Feb. 2019, pp. 1–8. DOI: 10.1109/ICACCP.2019.8882943. Accessed: Apr. 10, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8882943>.
- [18] A. K. A. Peñaloza, A. Balbinot, and R. Leborgne, "11 - AI application for load forecasting: A comparison of classical and deep learning methodologies," in *Monitoring and Control of Electrical Power Systems Using Machine Learning Techniques*, E. Barocio Espejo, F. R. Segundo Sevilla, and P. Korba, Eds., Elsevier, Jan. 2023, pp. 263–287, ISBN: 978-0-323-99904-5. DOI: 10.1016/B978-0-32-399904-5.00017-X. Accessed: Sep. 12, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B978032399904500017X>.
- [19] C. J. Sidey-Gibbons et al., "Predicting 180-day mortality for women with ovarian cancer using machine learning and patient-reported outcome data," in *Scientific Reports*, vol. 12, no. 1, p. 21269, Dec. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-022-22614-1. Accessed: Jun. 8, 2023. [Online]. Available: <https://www.nature.com/articles/s41598-022-22614-1>.
- [20] U.S. Department of Health and Human Services. Food and Drug Administration, "Guidance for Industry," en, 2009.
- [21] D. Denny, S. B. Nutakki, K. Alston, and M. Markman, "A review of patient self-reported symptom data with a focus on pain.," *Journal of Clinical Oncology*, vol. 32, no. 30_suppl, pp. 79–79, Oct. 2014, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/jco.2014.32.30_suppl.79. Accessed: Mar. 27, 2023. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/jco.2014.32.30_suppl.79.
- [22] D. Talati, "AI (Artificial Intelligence) in Daily Life," Sep. 2024. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.authorea.com/doi/full/10.36227/techrxiv.170751714.46556037?commit=5713f5c02d7f8f657a9b87381928070845a2906e>.
- [23] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *Sn Computer Science*, vol. 2, no. 5, p. 377, 2021, ISSN: 2662-995X. DOI: 10.1007/s42979-021-00765-8. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8274472/>.
- [24] K. Zhang, S. Liu, and M. Xiong, *Changes from Classical Statistics to Modern Statistics and Data Science*, arXiv:2211.03756 [stat], Oct. 2022. DOI: 10.48550/arXiv.2211.03756. Accessed: Sep. 4, 2025. [Online]. Available: <http://arxiv.org/abs/2211.03756>.
- [25] L. Faes, D. A. Sim, M. van Smeden, U. Held, P. M. Bossuyt, and L. M. Bachmann, "Artificial Intelligence and Statistics: Just the Old Wine in New Wineskins?" *Frontiers in Digital Health*, vol. 4, p. 833912, Jan. 2022, ISSN: 2673-253X. DOI: 10.3389/fdgth.2022.833912. Accessed: Sep. 4, 2025.

-
- [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8825497/>.
- [26] S. Friedrich et al., “Is there a role for statistics in artificial intelligence?” en, *Advances in Data Analysis and Classification*, vol. 16, no. 4, pp. 823–846, Dec. 2022, ISSN: 1862-5355. DOI: 10.1007/s11634-021-00455-6. Accessed: Sep. 4, 2025. [Online]. Available: <https://doi.org/10.1007/s11634-021-00455-6>.
- [27] D. Bzdok, N. Altman, and M. Krzywinski, “Statistics versus machine learning,” en, *Nature Methods*, vol. 15, no. 4, pp. 233–234, Apr. 2018, Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/nmeth.4642. Accessed: Aug. 6, 2025. [Online]. Available: <https://www.nature.com/articles/nmeth.4642>.
- [28] A. B. Rashid and M. A. K. Kausik, “AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications,” *Hybrid Advances*, vol. 7, p. 100277, Dec. 2024, ISSN: 2773-207X. DOI: 10.1016/j.hybadv.2024.100277. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2773207X24001386>.
- [29] E. J. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” eng, *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019, ISSN: 1546-170X. DOI: 10.1038/s41591-018-0300-7.
- [30] E. P. Adeghe, C. A. Okolo, O. T. Ojeyinka, E. P. Adeghe, C. A. Okolo, and O. T. Ojeyinka, “The role of big data in healthcare: A review of implications for patient outcomes and treatment personalization,” en, *World Journal of Biology Pharmacy and Health Sciences*, vol. 17, no. 3, pp. 198–204, 2024, Last Modified: 2024-05-25T12:31+05:30 Publisher: World Journal of Biology Pharmacy and Health Sciences, ISSN: 2582-5542, 2582-5542. DOI: 10.30574/wjbphs.2024.17.3.0133. Accessed: Sep. 4, 2025. [Online]. Available: <https://wjbphs.com/content/role-big-data-healthcare-review-implications-patient-outcomes-and-treatment-personalization>.
- [31] S. A. Alowais et al., “Revolutionizing healthcare: The role of artificial intelligence in clinical practice,” *BMC Medical Education*, vol. 23, no. 1, p. 689, Sep. 2023, ISSN: 1472-6920. DOI: 10.1186/s12909-023-04698-z. Accessed: Sep. 4, 2025. [Online]. Available: <https://doi.org/10.1186/s12909-023-04698-z>.
- [32] C. Kersting, M. Kneer, and A. Barzel, “Patient-relevant outcomes: What are we talking about? A scoping review to improve conceptual clarity,” *BMC Health Services Research*, vol. 20, no. 1, p. 596, Jun. 2020, ISSN: 1472-6963. DOI: 10.1186/s12913-020-05442-9. Accessed: Sep. 25, 2025. [Online]. Available: <https://doi.org/10.1186/s12913-020-05442-9>.
- [33] F. Jiang et al., “Artificial intelligence in healthcare: Past, present and future,” en, *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, Dec. 2017, ISSN: 2059-8688, 2059-8696. DOI: 10.1136/svn-2017-000101. Accessed: Nov. 28, 2022. [Online]. Available: <https://svn.bmj.com/lookup/doi/10.1136/svn-2017-000101>.

-
- [34] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," en, *Nature Medicine*, vol. 28, no. 1, pp. 31–38, Jan. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 1546-170X. DOI: 10.1038/s41591-021-01614-0. Accessed: Nov. 28, 2022. [Online]. Available: <https://www.nature.com/articles/s41591-021-01614-0>.
- [35] C. McGenity et al., "Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy," en, *npj Digital Medicine*, vol. 7, no. 1, p. 114, May 2024, Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-024-01106-8. Accessed: Sep. 17, 2025. [Online]. Available: <https://www.nature.com/articles/s41746-024-01106-8>.
- [36] S. Ganapathi et al., "Tackling bias in AI health datasets through the STAND-ING Together initiative," en, *Nature Medicine*, vol. 28, no. 11, pp. 2232–2233, Nov. 2022, ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-022-01987-w. Accessed: Nov. 26, 2025. [Online]. Available: <https://www.nature.com/articles/s41591-022-01987-w>.
- [37] S. Banerjee, P. Alsop, L. Jones, and R. N. Cardinal, "Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies," en, *Patterns*, vol. 3, no. 6, p. 100 506, Jun. 2022, ISSN: 26663899. DOI: 10.1016/j.patter.2022.100506. Accessed: Jul. 6, 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2666389922000988>.
- [38] R. M. Epstein and R. L. Street, "The Values and Value of Patient-Centered Care," en, *The Annals of Family Medicine*, vol. 9, no. 2, pp. 100–103, Mar. 2011, ISSN: 1544-1709, 1544-1717. DOI: 10.1370/afm.1239. Accessed: Jun. 28, 2023. [Online]. Available: <http://www.annfammed.org/cgi/doi/10.1370/afm.1239>.
- [39] K. D. Bertakis and R. Azari, "Patient-Centered Care is Associated with Decreased Health Care Utilization," en, *The Journal of the American Board of Family Medicine*, vol. 24, no. 3, pp. 229–239, May 2011, ISSN: 1557-2625, 1558-7118. DOI: 10.3122/jabfm.2011.03.100170. Accessed: Jun. 28, 2023. [Online]. Available: <http://www.jabfm.org/cgi/doi/10.3122/jabfm.2011.03.100170>.
- [40] A. Coulter and J. Oldham, "Person-centred care: What is it and how do we get there?" *Future Hospital Journal*, vol. 3, no. 2, pp. 114–116, Jun. 2016, ISSN: 2055-3323. DOI: 10.7861/futurehosp.3-2-114. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6465833/>.
- [41] A. Kwame and P. M. Petrucka, "A literature-based study of patient-centered care and communication in nurse-patient interactions: Barriers, facilitators, and the way forward," *BMC Nursing*, vol. 20, no. 1, p. 158, Sep. 2021, ISSN: 1472-6955. DOI: 10.1186/s12912-021-00684-2. Accessed: Sep. 4, 2025. [Online]. Available: <https://doi.org/10.1186/s12912-021-00684-2>.
- [42] N. Sartorius, "The Meanings of Health and its Promotion," *Croatian medical journal*, vol. 47, no. 4, pp. 662–664, Aug. 2006, ISSN: 0353-9504. Accessed: Sep. 23, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2080455/>.
-

-
- [43] R. D. Hays, C. D. Sherbourne, and R. Mazel, "User's Manual for the Medical Outcomes Study (MOS) Core Measures of Health-Related Quality of Life," en, Tech. Rep., Jan. 1995. Accessed: Sep. 23, 2025. [Online]. Available: https://www.rand.org/pubs/monograph_reports/MR162.html.
- [44] M. Bergner, R. A. Bobbitt, W. E. Pollard, D. P. Martin, and B. S. Gilson, "The sickness impact profile: Validation of a health status measure," eng, *Medical Care*, vol. 14, no. 1, pp. 57–67, Jan. 1976, ISSN: 0025-7079. DOI: 10.1097/00005650-197601000-00006.
- [45] N. K. Aaronson et al., "The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology," eng, *Journal of the National Cancer Institute*, vol. 85, no. 5, pp. 365–376, Mar. 1993, ISSN: 0027-8874. DOI: 10.1093/jnci/85.5.365.
- [46] D. F. Cella et al., "The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure.," *Journal of Clinical Oncology*, vol. 11, no. 3, pp. 570–579, Mar. 1993, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/JCO.1993.11.3.570. Accessed: Feb. 12, 2024.
- [47] T. Weldring and S. M. Smith, "Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs)," *Health Services Insights*, vol. 6, pp. 61–68, Aug. 2013, ISSN: 1178-6329. DOI: 10.4137/HSI.S11093. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4089835/>.
- [48] M. Calvert et al., "Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO extension," eng, *JAMA*, vol. 309, no. 8, pp. 814–822, Feb. 2013, ISSN: 1538-3598. DOI: 10.1001/jama.2013.879.
- [49] M. Calvert et al., "Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension," eng, *JAMA*, vol. 319, no. 5, pp. 483–494, Feb. 2018, ISSN: 1538-3598. DOI: 10.1001/jama.2017.21903.
- [50] The PROTEUS Consortium, *The Proteus Consortium*, en-US. Accessed: Sep. 24, 2025. [Online]. Available: <https://theproteusconsortium.org/>.
- [51] C. Coens et al., "International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: Recommendations of the SISAQOL Consortium," eng, *The Lancet. Oncology*, vol. 21, no. 2, e83–e96, Feb. 2020, ISSN: 1474-5488. DOI: 10.1016/S1470-2045(19)30790-9.
- [52] S. Cruz Rivera et al., "Ethical Considerations for the Inclusion of Patient-Reported Outcomes in Clinical Research: The PRO Ethics Guidelines," eng, *JAMA*, vol. 327, no. 19, pp. 1910–1919, May 2022, ISSN: 1538-3598. DOI: 10.1001/jama.2022.6421.

-
- [53] U. F. a. D. Administration, *Core Patient-Reported Outcomes in Cancer Clinical Trials*, en, Publisher: FDA, Oct. 2024. Accessed: Dec. 16, 2025. [Online]. Available: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/core-patient-reported-outcomes-cancer-clinical-trials>.
- [54] L. B. Kennedy and A. K. S. Salama, “A review of cancer immunotherapy toxicity,” eng, *CA: a cancer journal for clinicians*, vol. 70, no. 2, pp. 86–104, Mar. 2020, ISSN: 1542-4863. DOI: 10.3322/caac.21596.
- [55] C. Piccinin et al., “Flexibility in patient-reported outcome and health-related quality of life measurement: The EORTC Quality of Life Group measurement strategy,” eng, *European Journal of Cancer (Oxford, England: 1990)*, vol. 220, p. 115392, May 2025, ISSN: 1879-0852. DOI: 10.1016/j.ejca.2025.115392.
- [56] K. A. Webster et al., “Development of a Functional Assessment of Chronic Illness Therapy item library and primary symptom list for the assessment of patient-reported adverse events associated with immune checkpoint modulators,” eng, *Journal of Cancer Metastasis and Treatment*, vol. 6, p. 8, 2020, ISSN: 2394-4722. DOI: 10.20517/2394-4722.2019.38.
- [57] C. Piccinin et al., “Recommendations on the use of item libraries for patient-reported outcome measurement in oncology trials: Findings from an international, multidisciplinary working group,” eng, *The Lancet. Oncology*, vol. 24, no. 2, e86–e95, Feb. 2023, ISSN: 1474-5488. DOI: 10.1016/S1470-2045(22)00654-4.
- [58] E. Basch et al., “Development of the National Cancer Institute’s Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE),” *JNCI Journal of the National Cancer Institute*, vol. 106, no. 9, dju244, Sep. 2014, ISSN: 0027-8874. DOI: 10.1093/jnci/dju244. Accessed: Sep. 17, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4200059/>.
- [59] A. Trotti et al., “Common toxicity criteria: Version 2.0. an improved reference for grading the acute effects of cancer treatment: Impact on radiotherapy,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 47, no. 1, pp. 13–47, Apr. 2000, ISSN: 0360-3016. DOI: 10.1016/S0360-3016(99)00559-3. Accessed: Aug. 11, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360301699005593>.
- [60] E. Basch et al., “Long-term toxicity monitoring via electronic patient-reported outcomes in patients receiving chemotherapy,” eng, *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 25, no. 34, pp. 5374–5380, Dec. 2007, ISSN: 1527-7755. DOI: 10.1200/JCO.2007.11.2243.
- [61] P. Holch et al., “Asking the right questions to get the right answers: Using cognitive interviews to review the acceptability, comprehension and clinical meaningfulness of patient self-report adverse event items in oncology patients,” eng, *Acta Oncologica (Stockholm, Sweden)*, vol. 55, no. 9-10, pp. 1220–1226, 2016, ISSN: 1651-226X. DOI: 10.1080/0284186X.2016.1213878.
-

-
- [62] O. L. Aiyegbusi et al., “Key considerations to reduce or address respondent burden in patient-reported outcome (PRO) data collection,” en, *Nature Communications*, vol. 13, no. 1, p. 6026, Oct. 2022, Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-022-33826-4. Accessed: Dec. 16, 2025. [Online]. Available: <https://www.nature.com/articles/s41467-022-33826-4>.
- [63] A. Awad et al., “Connected healthcare: Improving patient care using digital health technologies,” en, *Advanced Drug Delivery Reviews*, vol. 178, p. 113958, Nov. 2021, ISSN: 0169-409X. DOI: 10.1016/j.addr.2021.113958. Accessed: Nov. 28, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169409X21003513>.
- [64] J. Y. Yu, T. Goldberg, N. Lao, B. M. Feldman, and Y. I. Goh, “Electronic forms for patient reported outcome measures (PROMs) are an effective, time-efficient, and cost-minimizing alternative to paper forms,” *Pediatric Rheumatology Online Journal*, vol. 19, p. 67, May 2021, ISSN: 1546-0096. DOI: 10.1186/s12969-021-00551-z. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8091685/>.
- [65] J. Lehmann, T. Cofala, M. Tschuggnall, J. M. Giesinger, G. Rumpold, and B. Holzner, “Machine learning in oncology—Perspectives in patient-reported outcome research,” en, *Der Onkologe*, vol. 27, no. S2, pp. 150–155, Nov. 2021, ISSN: 0947-8965, 1433-0415. DOI: 10.1007/s00761-021-00916-9. Accessed: Aug. 29, 2022. [Online]. Available: <https://link.springer.com/10.1007/s00761-021-00916-9>.
- [66] G. Velikova et al., “Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial,” eng, *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 22, no. 4, pp. 714–724, Feb. 2004, ISSN: 0732-183X. DOI: 10.1200/JCO.2004.06.078.
- [67] E. Basch et al., “Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial,” eng, *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 34, no. 6, pp. 557–565, Feb. 2016, ISSN: 1527-7755. DOI: 10.1200/JCO.2015.63.0830.
- [68] E. Basch et al., “Effect of Electronic Symptom Monitoring on Patient-Reported Outcomes Among Patients With Metastatic Cancer: A Randomized Clinical Trial,” eng, *JAMA*, vol. 327, no. 24, pp. 2413–2422, Jun. 2022, ISSN: 1538-3598. DOI: 10.1001/jama.2022.9265.
- [69] F. Denis et al., “Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer,” eng, *JAMA*, vol. 321, no. 3, pp. 306–307, Jan. 2019, ISSN: 1538-3598. DOI: 10.1001/jama.2018.18085.
- [70] K. Absolom et al., “Phase III Randomized Controlled Trial of eRAPID: eHealth Intervention During Chemotherapy,” *Journal of Clinical Oncology*, vol. 39, no. 7, pp. 734–747, Mar. 2021, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/JCO.20.02015. Accessed: Aug. 29, 2022.

-
- [71] E. C. Nelson, E. Eftimovska, C. Lind, A. Hager, J. H. Wasson, and S. Lindblad, "Patient reported outcome measures in practice," en, *BMJ*, g7818, Feb. 2015, ISSN: 1756-1833. DOI: 10.1136/bmj.g7818. Accessed: Jun. 28, 2023. [Online]. Available: <https://www.bmj.com/lookup/doi/10.1136/bmj.g7818>.
- [72] S. C. Rivera et al., "Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies," English, *The Lancet Digital Health*, vol. 5, no. 3, e168–e173, Mar. 2023, Publisher: Elsevier, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(22)00252-7. Accessed: Jul. 3, 2023. [Online]. Available: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(22\)00252-7/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00252-7/fulltext).
- [73] D. Verma, K. Bach, and P. J. Mork, "Application of Machine Learning Methods on Patient Reported Outcome Measurements for Predicting Outcomes: A Literature Review," en, *Informatics*, vol. 8, no. 3, p. 56, Sep. 2021, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-9709. DOI: 10.3390/informatics8030056. Accessed: Jun. 12, 2023. [Online]. Available: <https://www.mdpi.com/2227-9709/8/3/56>.
- [74] F. J. Pearce, S. Cruz Rivera, X. Liu, E. Manna, A. K. Denniston, and M. J. Calvert, "The role of patient-reported outcome measures in trials of artificial intelligence health technologies: A systematic evaluation of ClinicalTrials.gov records (1997–2022)," *The Lancet Digital Health*, vol. 5, no. 3, e160–e167, Mar. 2023, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(22)00249-7. Accessed: Sep. 10, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2589750022002497>.
- [75] A. Katakam et al., "Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty," en, *Journal of Orthopaedic Research*, vol. 40, no. 4, pp. 808–815, 2022, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jor.25125>, ISSN: 1554-527X. DOI: 10.1002/jor.25125. Accessed: Jan. 15, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jor.25125>.
- [76] M. Tschuggnall, V. Grote, M. Pirchl, B. Holzner, G. Rumpold, and M. J. Fischer, "Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures," en, *Informatics in Medicine Unlocked*, vol. 24, p. 100598, Jan. 2021, ISSN: 2352-9148. DOI: 10.1016/j.imu.2021.100598. Accessed: Aug. 23, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914821000885>.
- [77] S. Jayakumar et al., "Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: A meta-research study," en, *npj Digital Medicine*, vol. 5, no. 1, pp. 1–13, Jan. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-021-00544-y. Accessed: Jan. 3, 2024. [Online]. Available: <https://www.nature.com/articles/s41746-021-00544-y>.

-
- [78] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, “Responsible AI for Digital Health: A Synthesis and a Research Agenda,” en, *Information Systems Frontiers*, vol. 25, no. 6, pp. 2139–2157, Dec. 2023, ISSN: 1572-9419. DOI: 10.1007/s10796-021-10146-4. Accessed: Jan. 3, 2024. [Online]. Available: <https://doi.org/10.1007/s10796-021-10146-4>.
- [79] B. Khan et al., “Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector,” *Biomedical Materials & Devices (New York, N.y.)*, pp. 1–8, Feb. 2023, ISSN: 2731-4812. DOI: 10.1007/s44174-023-00063-2. Accessed: Jan. 3, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9908503/>.
- [80] T. Engstrom et al., “Patient reported outcome measure domains and tools used among adolescents and young adults with cancer: A scoping review,” *Critical Reviews in Oncology/Hematology*, vol. 181, p. 103867, Jan. 2023, ISSN: 1040-8428. DOI: 10.1016/j.critrevonc.2022.103867. Accessed: Jan. 3, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1040842822002918>.
- [81] P. Jayakumar et al., “Digital Phenotyping and Patient-Generated Health Data for Outcome Measurement in Surgical Care: A Scoping Review,” en, *Journal of Personalized Medicine*, vol. 10, no. 4, p. 282, Dec. 2020, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2075-4426. DOI: 10.3390/jpm10040282. Accessed: Jan. 3, 2024. [Online]. Available: <https://www.mdpi.com/2075-4426/10/4/282>.
- [82] M. D. J. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares, “Guidance for conducting systematic scoping reviews,” eng, *International Journal of Evidence-Based Healthcare*, vol. 13, no. 3, pp. 141–146, Sep. 2015, ISSN: 1744-1609. DOI: 10.1097/XEB.0000000000000050.
- [83] Z. Wojcik, V. Dimitrova, L. Warrington, G. Velikova, and K. Absolom, “Predicting patient outcomes from self-reported symptom data using Artificial intelligence: A scoping review protocol,” en, Sep. 2023, Publisher: OSF. Accessed: Apr. 12, 2024. [Online]. Available: <https://osf.io/2rpf6/>.
- [84] A. C. Tricco et al., “PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation,” en, *Annals of Internal Medicine*, vol. 169, no. 7, pp. 467–473, Oct. 2018, ISSN: 0003-4819, 1539-3704. DOI: 10.7326/M18-0850. Accessed: Apr. 12, 2024. [Online]. Available: <https://www.acpjournals.org/doi/10.7326/M18-0850>.
- [85] E. M. Roos, H. P. Roos, L. S. Lohmander, C. Ekdahl, and B. D. Beynnon, “Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a Self-Administered Outcome Measure,” *Journal of Orthopaedic & Sports Physical Therapy*, vol. 28, no. 2, pp. 88–96, Aug. 1998, Publisher: Journal of Orthopaedic & Sports Physical Therapy, ISSN: 0190-6011. DOI: 10.2519/jospt.1998.28.2.88. Accessed: Feb. 21, 2024. [Online]. Available: <https://www.jospt.org/doi/abs/10.2519/jospt.1998.28.2.88>.
- [86] R. Rabin and F. de Charro, “EQ-5D: A measure of health status from the EuroQol Group,” eng, *Annals of Medicine*, vol. 33, no. 5, pp. 337–343, Jul. 2001, ISSN: 0785-3890. DOI: 10.3109/07853890109002087.

-
- [87] S. P. Koyyada and T. P. Singh, “A multi stage approach to handle class imbalance:An ensemble method,” en, *Procedia Computer Science*, vol. 218, pp. 2666–2674, 2023, ISSN: 18770509. DOI: 10.1016/j.procs.2023.01.239. Accessed: Dec. 11, 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1877050923002399>.
- [88] Y. Wang et al., “Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes,” eng, *Proceedings. International Database Engineering and Applications Symposium*, vol. 2021, pp. 273–279, Jul. 2021, ISSN: 1098-8068. DOI: 10.1145/3472163.3472177.
- [89] M. Kalweit et al., “Personalized prediction of disease activity in patients with rheumatoid arthritis using an adaptive deep neural network,” eng, *PloS One*, vol. 16, no. 6, e0252289, 2021, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0252289.
- [90] R. C. Grant et al., “Machine Learning–Based Early Warning Systems for Acute Care Utilization During Systemic Therapy for Cancer,” EN, *Journal of the National Comprehensive Cancer Network*, vol. 21, no. 10, 1029–1037.e21, Oct. 2023, Publisher: National Comprehensive Cancer Network Section: Journal of the National Comprehensive Cancer Network, ISSN: 1540-1405, 1540-1413. DOI: 10.6004/jnccn.2023.7046. Accessed: Jan. 15, 2024. [Online]. Available: <https://jnccn.org/view/journals/jnccn/21/10/article-p1029.xml>.
- [91] C. Klemm et al., “The utility of machine learning algorithms for the prediction of patient-reported outcome measures following primary hip and knee total joint arthroplasty,” en, *Archives of Orthopaedic and Trauma Surgery*, vol. 143, no. 4, pp. 2235–2245, Apr. 2023, ISSN: 1434-3916. DOI: 10.1007/s00402-022-04526-x. Accessed: Jul. 2, 2023. [Online]. Available: <https://doi.org/10.1007/s00402-022-04526-x>.
- [92] L. G. Tennenhouse, R. A. Marrie, C. N. Bernstein, and L. M. Lix, “Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease,” *Journal of Psychosomatic Research*, vol. 134, p. 110126, Jul. 2020, ISSN: 0022-3999. DOI: 10.1016/j.jpsychores.2020.110126. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0022399919308785>.
- [93] S. Zhang, J. Y. Chen, H. N. Pang, N. N. Lo, S. J. Yeo, and M. H. L. Liow, “Development and internal validation of machine learning algorithms to predict patient satisfaction after total hip arthroplasty,” *Arthroplasty*, vol. 3, no. 1, p. 33, Sep. 2021, ISSN: 2524-7948. DOI: 10.1186/s42836-021-00087-3. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1186/s42836-021-00087-3>.
- [94] S. Zhang, B. P. H. Lau, Y. H. Ng, X. Wang, and W. Chua, “Machine learning algorithms do not outperform preoperative thresholds in predicting clinically meaningful improvements after total knee arthroplasty,” en, *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 30, no. 8, pp. 2624–2630, Aug. 2022, ISSN: 1433-7347. DOI: 10.1007/s00167-021-06642-4. Accessed: Jan. 15,

-
2024. [Online]. Available: <https://doi.org/10.1007/s00167-021-06642-4>.
- [95] D. Verma, K. Bach, and P. J. Mork, “Using Automated Feature Selection for Building Case-Based Reasoning Systems: An Example from Patient-Reported Outcome Measurements,” in *Artificial Intelligence XXXVIII: 41st SGA International Conference on Artificial Intelligence, AI 2021, Cambridge, UK, December 14–16, 2021, Proceedings*, Berlin, Heidelberg: Springer-Verlag, Dec. 2021, pp. 282–295, ISBN: 978-3-030-91099-0. DOI: 10.1007/978-3-030-91100-3_23. Accessed: Jan. 14, 2024. [Online]. Available: https://doi.org/10.1007/978-3-030-91100-3_23.
- [96] A. A. Ø. Buus, F. W. Udsen, B. Laugesen, A. El-Galaly, M. Laursen, and O. K. Hejlesen, “Patient-Reported Outcomes for Function and Pain in Total Knee Arthroplasty Patients,” eng, *Nursing Research*, vol. 71, no. 5, E39–E47, Oct. 2022, ISSN: 1538-9847. DOI: 10.1097/NNR.0000000000000602.
- [97] V. E. Staartjes, M. P. de Wispelaere, W. P. Vandertop, and M. L. Schröder, “Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: Feasibility of center-specific modeling,” eng, *The Spine Journal: Official Journal of the North American Spine Society*, vol. 19, no. 5, pp. 853–861, May 2019, ISSN: 1878-1632. DOI: 10.1016/j.spinee.2018.11.009.
- [98] M. Huber, C. Kurz, and R. Leidl, “Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning,” *BMC Medical Informatics and Decision Making*, vol. 19, p. 3, Jan. 2019, ISSN: 1472-6947. DOI: 10.1186/s12911-018-0731-6. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6325823/>.
- [99] C. Xu, I. M. Subbiah, S.-C. Lu, A. Pfob, and C. Sidey-Gibbons, “Machine learning models for 180-day mortality prediction of patients with advanced cancer using patient-reported symptom data,” en, *Quality of Life Research*, Oct. 2022, ISSN: 0962-9343, 1573-2649. DOI: 10.1007/s11136-022-03284-y. Accessed: Nov. 1, 2022. [Online]. Available: <https://link.springer.com/10.1007/s11136-022-03284-y>.
- [100] A. H. S. Harris, A. C. Kuo, T. R. Bowe, L. Manfredi, N. F. Lalani, and N. J. Giori, “Can Machine Learning Methods Produce Accurate and Easy-to-Use Preoperative Prediction Models of One-Year Improvements in Pain and Functioning After Knee Arthroplasty?” English, *The Journal of Arthroplasty*, vol. 36, no. 1, 112–117.e6, Jan. 2021, Publisher: Elsevier, ISSN: 0883-5403, 1532-8406. DOI: 10.1016/j.arth.2020.07.026. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.arthroplastyjournal.org/article/S0883-5403\(20\)30783-X/fulltext](https://www.arthroplastyjournal.org/article/S0883-5403(20)30783-X/fulltext).
- [101] A. Siccoli, M. P. d. Wispelaere, M. L. Schröder, and V. E. Staartjes, “Machine learning-based preoperative predictive analytics for lumbar spinal stenosis,” EN, *Neurosurgical Focus*, vol. 46, no. 5, E5, May 2019, Publisher: American Association of Neurological Surgeons Section: Neurosurgical Focus, ISSN: 1092-0684. DOI: 10.3171/2019.2.FOCUS18723. Accessed: Jan. 15, 2024. [On-

-
- line]. Available: <https://thejns.org/focus/view/journals/neurosurg-focus/46/5/article-pE5.xml>.
- [102] A. Pfob, B. J. Mehrara, J. A. Nelson, E. G. Wilkins, A. L. Pusic, and C. Sidey-Gibbons, "Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPIRED-001)," English, *The Breast*, vol. 60, pp. 111–122, Dec. 2021, Publisher: Elsevier, ISSN: 0960-9776, 1532-3080. DOI: 10.1016/j.breast.2021.09.009. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.thebreastonline.com/article/S0960-9776\(21\)00466-5/fulltext](https://www.thebreastonline.com/article/S0960-9776(21)00466-5/fulltext).
- [103] C. F. Pedersen, M. Ø. Andersen, L. Y. Carreon, and S. Eiskjær, "Applied Machine Learning for Spine Surgeons: Predicting Outcome for Patients Undergoing Treatment for Lumbar Disc Herniation Using PRO Data," *Global Spine Journal*, vol. 12, no. 5, pp. 866–876, Jun. 2022, ISSN: 2192-5682. DOI: 10.1177/2192568220967643. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9344505/>.
- [104] C. B. Josephson et al., "Predicting postoperative epilepsy surgery satisfaction in adults using the 19-item Epilepsy Surgery Satisfaction Questionnaire and machine learning," en, *Epilepsia*, vol. 62, no. 9, pp. 2103–2112, 2021, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.16992>, ISSN: 1528-1167. DOI: 10.1111/epi.16992. Accessed: Jan. 15, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.16992>.
- [105] P. N. Ramkumar, J. M. Karnuta, H. S. Haeberle, S. A. Rodeo, B. U. Nwachukwu, and R. J. Williams, "Effect of Preoperative Imaging and Patient Factors on Clinically Meaningful Outcomes and Quality of Life After Osteochondral Allograft Transplantation: A Machine Learning Analysis of Cartilage Defects of the Knee," en, *The American Journal of Sports Medicine*, vol. 49, no. 8, pp. 2177–2186, Jul. 2021, Publisher: SAGE Publications Inc STM, ISSN: 0363-5465. DOI: 10.1177/03635465211015179. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1177/03635465211015179>.
- [106] J. S. Munn et al., "Logistic Regression and Machine Learning Models Cannot Discriminate Between Satisfied and Dissatisfied Total Knee Arthroplasty Patients," English, *The Journal of Arthroplasty*, vol. 37, no. 2, pp. 267–273, Feb. 2022, Publisher: Elsevier, ISSN: 0883-5403, 1532-8406. DOI: 10.1016/j.arth.2021.10.017. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.arthroplastyjournal.org/article/S0883-5403\(21\)00822-6/fulltext](https://www.arthroplastyjournal.org/article/S0883-5403(21)00822-6/fulltext).
- [107] K. M. Kober et al., "Prediction of evening fatigue severity in outpatients receiving chemotherapy: Less may be more," *Fatigue : biomedicine, health & behavior*, vol. 9, no. 1, pp. 14–32, 2021, ISSN: 2164-1846. DOI: 10.1080/21641846.2021.1885119. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8262130/>.
- [108] W. M. Durand et al., "Artificial Intelligence Models Predict Operative Versus Nonoperative Management of Patients with Adult Spinal Deformity with 86% Accuracy," *World Neurosurgery*, vol. 141, e239–e253, Sep. 2020, ISSN: 1878-8750. DOI: 10.1016/j.wneu.2020.05.099. Accessed: Jan. 15, 2024.
-

-
- [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1878875020310706>.
- [109] C. J. Harrison, L. Geoghegan, C. J. Sidey-Gibbons, P. H. C. Stirling, J. E. McEachan, and J. N. Rodrigues, "Developing Machine Learning Algorithms to Support Patient-centered, Value-based Carpal Tunnel Decompression Surgery," *Plastic and Reconstructive Surgery Global Open*, vol. 10, no. 4, e4279, Apr. 2022, ISSN: 2169-7574. DOI: 10.1097/GOX.0000000000004279. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9015194/>.
- [110] V. Kumar et al., "What Is the Accuracy of Three Different Machine Learning Techniques to Predict Clinical Outcomes After Shoulder Arthroplasty?" en-US, *Clinical Orthopaedics and Related Research®*, vol. 478, no. 10, p. 2351, Oct. 2020, ISSN: 0009-921X. DOI: 10.1097/CORR.0000000000001263. Accessed: Jan. 15, 2024. [Online]. Available: https://journals.lww.com/clinorthop/fulltext/2020/10000/what_is_the_accuracy_of_three_different_machine.26.aspx.
- [111] S. Wshah, C. Skalka, and M. Price, "Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach," *JMIR Mental Health*, vol. 6, no. 7, e13946, Jul. 2019, ISSN: 2368-7959. DOI: 10.2196/13946. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6681635/>.
- [112] Y. Lu et al., "Machine-learning model successfully predicts patients at risk for prolonged postoperative opioid use following elective knee arthroscopy," eng, *Knee surgery, sports traumatology, arthroscopy: official journal of the ESSKA*, vol. 30, no. 3, pp. 762–772, Mar. 2022, ISSN: 1433-7347. DOI: 10.1007/s00167-020-06421-7.
- [113] K. N. Kunze, A. V. Karhade, A. J. Sadauskas, J. H. Schwab, and B. R. Levine, "Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty," *The Journal of Arthroplasty*, vol. 35, no. 8, pp. 2119–2123, Aug. 2020, ISSN: 0883-5403. DOI: 10.1016/j.arth.2020.03.019. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0883540320302679>.
- [114] N. L. Loos et al., "Machine Learning Can be Used to Predict Function but Not Pain After Surgery for Thumb Carpometacarpal Osteoarthritis," eng, *Clinical Orthopaedics and Related Research*, vol. 480, no. 7, pp. 1271–1284, Jul. 2022, ISSN: 1528-1132. DOI: 10.1097/CORR.0000000000002105.
- [115] O. Shipston-Sharman, S. Popkirov, C. H. Hansen, J. Stone, and A. Carson, "Prognosis in functional and recognised pathophysiological neurological disorders - a shared basis," eng, *Journal of Psychosomatic Research*, vol. 152, p. 110681, Nov. 2021, ISSN: 1879-1360. DOI: 10.1016/j.jpsychores.2021.110681.

-
- [116] B. K. Tan et al., “Effect of symptom-based risk stratification on the costs of managing patients with chronic rhinosinusitis symptoms,” en, *International Forum of Allergy & Rhinology*, vol. 3, no. 11, pp. 933–940, 2013, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/alr.21208>, ISSN: 2042-6984. DOI: 10.1002/alr.21208. Accessed: Jan. 15, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/alr.21208>.
- [117] A. A. Rahman et al., “Perceived Stress Analysis of Undergraduate Students during COVID-19: A Machine Learning Approach,” in *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*, ISSN: 2158-8481, Jun. 2022, pp. 1129–1134. DOI: 10.1109/MELECON53508.2022.9843081. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9843081>.
- [118] E. S. Lee, “Exploring the Performance of Stacking Classifier to Predict Depression Among the Elderly,” in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Aug. 2017, pp. 13–20. DOI: 10.1109/ICHI.2017.95. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/8031126>.
- [119] Z. Nowinka, M. A. Alagha, K. Mahmoud, and G. G. Jones, “Predicting Depression in Patients With Knee Osteoarthritis Using Machine Learning: Model Development and Validation Study,” eng, *JMIR formative research*, vol. 6, no. 9, e36130, Sep. 2022, ISSN: 2561-326X. DOI: 10.2196/36130.
- [120] F. S. Arkin, G. Aras, and E. Dogu, “Comparison of Artificial Neural Networks and Logistic Regression for 30-days Survival Prediction of Cancer Patients,” eng, *Acta informatica medica: AIM: journal of the Society for Medical Informatics of Bosnia & Herzegovina: casopis Drustva za medicinsku informatiku BiH*, vol. 28, no. 2, pp. 108–113, Jun. 2020, ISSN: 0353-8109. DOI: 10.5455/aim.2020.28.108–113.
- [121] L. Hoogendam et al., “Predicting Clinically Relevant Patient-Reported Symptom Improvement After Carpal Tunnel Release: A Machine Learning Approach,” eng, *Neurosurgery*, vol. 90, no. 1, pp. 106–113, Jan. 2022, ISSN: 1524-4040. DOI: 10.1227/NEU.0000000000001749.
- [122] E. J. Camp et al., “Supervised machine learning to predict reduced depression severity in people with epilepsy through epilepsy self-management intervention,” English, *Epilepsy & Behavior*, vol. 127, Feb. 2022, Publisher: Elsevier, ISSN: 1525-5050, 1525-5069. DOI: 10.1016/j.yebeh.2021.108548. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.epilepsybehavior.com/article/S1525-5050\(21\)00809-X/fulltext](https://www.epilepsybehavior.com/article/S1525-5050(21)00809-X/fulltext).
- [123] K. J. Wardenaar et al., “Common and specific determinants of 9-year depression and anxiety course-trajectories: A machine-learning investigation in the Netherlands Study of Depression and Anxiety (NESDA).,” *Journal of Affective Disorders*, vol. 293, pp. 295–304, Oct. 2021, ISSN: 0165-0327. DOI: 10.1016/j.jad.2021.06.029. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0165032721006157>.
-

-
- [124] K. N. Kunze, E. M. Polce, B. U. Nwachukwu, J. Chahla, and S. J. Nho, "Development and Internal Validation of Supervised Machine Learning Algorithms for Predicting Clinically Significant Functional Improvement in a Mixed Population of Primary Hip Arthroscopy," English, *Arthroscopy*, vol. 37, no. 5, pp. 1488–1497, May 2021, Publisher: Elsevier, ISSN: 0749-8063, 1526-3231. DOI: 10.1016/j.arthro.2021.01.005. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.arthroscopyjournal.org/article/S0749-8063\(21\)00016-5/fulltext](https://www.arthroscopyjournal.org/article/S0749-8063(21)00016-5/fulltext).
- [125] D. L. Smith and P. Held, "Moving toward precision PTSD treatment: Predicting veterans' intensive PTSD treatment response using continuously updating machine learning models," eng, *Psychological Medicine*, vol. 53, no. 12, pp. 5500–5509, Sep. 2023, ISSN: 1469-8978. DOI: 10.1017/S0033291722002689.
- [126] C. Bone et al., "Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data," en, *The Lancet Digital Health*, vol. 3, no. 4, e231–e240, Apr. 2021, ISSN: 25897500. DOI: 10.1016/S2589-7500(21)00018-2. Accessed: Jan. 15, 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2589750021000182>.
- [127] C. O'Driscoll et al., "The importance of transdiagnostic symptom level assessment to understanding prognosis for depressed adults: Analysis of data from six randomised control trials," *BMC Medicine*, vol. 19, no. 1, p. 109, May 2021, ISSN: 1741-7015. DOI: 10.1186/s12916-021-01971-0. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1186/s12916-021-01971-0>.
- [128] H.-Y. Shi, J.-T. Tsai, Y.-M. Chen, R. Culbertson, H.-T. Chang, and M.-F. Hou, "Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models," eng, *Breast Cancer Research and Treatment*, vol. 135, no. 1, pp. 221–229, Aug. 2012, ISSN: 1573-7217. DOI: 10.1007/s10549-012-2174-6.
- [129] Y.-H. Pua et al., "Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty," en, *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 28, no. 10, pp. 3207–3216, Oct. 2020, ISSN: 1433-7347. DOI: 10.1007/s00167-019-05822-7. Accessed: Aug. 23, 2022. [Online]. Available: <https://doi.org/10.1007/s00167-019-05822-7>.
- [130] B. Langenberger, "Who will stay a little longer? Predicting length of stay in hip and knee arthroplasty patients using machine learning," *Intelligence-Based Medicine*, vol. 8, p. 100111, Jan. 2023, ISSN: 2666-5212. DOI: 10.1016/j.ibmed.2023.100111. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S266652122300025X>.
- [131] H. Hasannejadasl et al., "A comparison of machine learning models for predicting urinary incontinence in men with localized prostate cancer," *Frontiers in Oncology*, vol. 13, p. 1168219, Apr. 2023, ISSN: 2234-943X. DOI: 10.3389/fonc.2023.1168219. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10130634/>.

-
- [132] C. Xu et al., “Enhanced Surgical Decision-Making Tools in Breast Cancer: Predicting 2-Year Postoperative Physical, Sexual, and Psychosocial Well-Being following Mastectomy and Breast Reconstruction (INSPIRED 004),” eng, *Annals of Surgical Oncology*, vol. 30, no. 12, pp. 7046–7059, Nov. 2023, ISSN: 1534-4681. DOI: 10.1245/s10434-023-13971-w. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1245/s10434-023-13971-w>.
- [133] J. Tian et al., “Machine learning prognosis model based on patient-reported outcomes for chronic heart failure patients after discharge,” *Health and Quality of Life Outcomes*, vol. 21, no. 1, p. 31, Mar. 2023, ISSN: 1477-7525. DOI: 10.1186/s12955-023-02109-x. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1186/s12955-023-02109-x>.
- [134] C. Park et al., “Which supervised machine learning algorithm can best predict achievement of minimum clinically important difference in neck pain after surgery in patients with cervical myelopathy? A QOD study,” EN, *Neurosurgical Focus*, vol. 54, no. 6, E5, Jun. 2023, Publisher: American Association of Neurological Surgeons Section: Neurosurgical Focus, ISSN: 1092-0684. DOI: 10.3171/2023.3.FOCUS2372. Accessed: Jan. 15, 2024. [Online]. Available: <https://thejns.org/focus/view/journals/neurosurg-focus/54/6/article-pE5.xml>.
- [135] B. Langenberger, D. Schrednitzki, A. M. Halder, R. Busse, and C. M. Pross, “Predicting whether patients will achieve minimal clinically important differences following hip or knee arthroplasty,” *Bone & Joint Research*, vol. 12, no. 9, pp. 512–521, Sep. 2023, ISSN: 2046-3758. DOI: 10.1302/2046-3758.129.BJR-2023-0070.R2. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10471446/>.
- [136] J. R. Curtis et al., “Machine Learning Applied to Patient-Reported Outcomes to Classify Physician-Derived Measures of Rheumatoid Arthritis Disease Activity,” eng, *ACR open rheumatology*, vol. 4, no. 12, pp. 995–1003, Dec. 2022, ISSN: 2578-5745. DOI: 10.1002/acr2.11499.
- [137] S. Iivanainen, J. Ekstrom, H. Virtanen, V. V. Kataja, and J. P. Koivunen, “Electronic patient-reported outcomes and machine learning in predicting immune-related adverse events of immune checkpoint inhibitor therapies,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 205, Jun. 2021, ISSN: 1472-6947. DOI: 10.1186/s12911-021-01564-0. Accessed: Jul. 2, 2023. [Online]. Available: <https://doi.org/10.1186/s12911-021-01564-0>.
- [138] C. W. Noel et al., “Development and Validation of a Machine Learning Algorithm Predicting Emergency Department Use and Unplanned Hospitalization in Patients With Head and Neck Cancer,” eng, *JAMA otolaryngology–head & neck surgery*, vol. 148, no. 8, pp. 764–772, Aug. 2022, ISSN: 2168-619X. DOI: 10.1001/jamaoto.2022.1629.
- [139] F. P. Chmiel, D. K. Burns, J. B. Pickering, A. Blythin, T. M. Wilkinson, and M. J. Boniface, “Prediction of Chronic Obstructive Pulmonary Disease Exacerbation Events by Using Patient Self-reported Data in a Digital Health App: Statistical Evaluation and Machine Learning Approach,” eng, *JMIR*
-

-
- medical informatics*, vol. 10, no. 3, e26499, Mar. 2022, ISSN: 2291-9694. DOI: 10.2196/26499.
- [140] H. N. Ziobrowski et al., “Development and Validation of a Model to Predict Posttraumatic Stress Disorder and Major Depression After a Motor Vehicle Collision,” eng, *JAMA psychiatry*, vol. 78, no. 11, pp. 1228–1237, Nov. 2021, ISSN: 2168-6238. DOI: 10.1001/jamapsychiatry.2021.2427.
- [141] P. Annapureddy et al., “Predicting PTSD Severity in Veterans from Self-reports for Early Intervention: A Machine Learning Approach,” in *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, Aug. 2020, pp. 201–208. DOI: 10.1109/IRI49571.2020.00036. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9191375>.
- [142] M. Zhong, V. Van Zoest, A. M. Bilal, F. Papadopoulos, and G. Castellano, “Unimodal vs. Multimodal Prediction of Antenatal Depression from Smartphone-based Survey Data in a Longitudinal Study,” in *Proceedings of the 2022 International Conference on Multimodal Interaction*, ser. ICMI ’22, New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 455–467, ISBN: 978-1-4503-9390-4. DOI: 10.1145/3536221.3556605. Accessed: Apr. 11, 2024. [Online]. Available: <https://dl.acm.org/doi/10.1145/3536221.3556605>.
- [143] L. S. Canas et al., “Early detection of COVID-19 in the UK using self-reported symptoms: A large-scale, prospective, epidemiological surveillance study,” English, *The Lancet Digital Health*, vol. 3, no. 9, e587–e598, Sep. 2021, Publisher: Elsevier, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(21)00131-X. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(21\)00131-X/fulltext](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00131-X/fulltext).
- [144] R. Sutradhar and L. Barbera, “Comparing an Artificial Neural Network to Logistic Regression for Predicting ED Visit Risk Among Patients With Cancer: A Population-Based Cohort Study,” English, *Journal of Pain and Symptom Management*, vol. 60, no. 1, pp. 1–9, Jul. 2020, Publisher: Elsevier, ISSN: 0885-3924, 1873-6513. DOI: 10.1016/j.jpainsymman.2020.02.010. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.jpmsjournal.com/article/S0885-3924\(20\)30096-8/fulltext](https://www.jpmsjournal.com/article/S0885-3924(20)30096-8/fulltext).
- [145] N. Goldstein et al., “Exploring Early Pre-Symptomatic Detection of Influenza Using Continuous Monitoring of Advanced Physiological Parameters during a Randomized Controlled Trial,” *Journal of Clinical Medicine*, vol. 10, no. 21, p. 5202, Nov. 2021, ISSN: 2077-0383. DOI: 10.3390/jcm10215202. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8584386/>.
- [146] Q. A. Rahman, T. Janmohamed, H. Clarke, P. Ritvo, J. Heffernan, and J. Katz, “Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods,” *JMIR Medical Informatics*, vol. 7, no. 4, e15601, Nov. 2019, ISSN: 2291-9694. DOI: 10.2196/15601. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6913759/>.
-

-
- [147] D. Verma, D. Jansen, K. Bach, M. Poel, P. J. Mork, and W. O. N. d'Hollosy, "Exploratory application of machine learning methods on patient reported data in the development of supervised models for predicting outcomes," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 227, Sep. 2022, ISSN: 1472-6947. DOI: 10.1186/s12911-022-01973-9. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1186/s12911-022-01973-9>.
- [148] M. G. Crowson et al., "Predicting Postoperative Cochlear Implant Performance Using Supervised Machine Learning," en-US, *Otology & Neurotology*, vol. 41, no. 8, e1013, Sep. 2020, ISSN: 1531-7129. DOI: 10.1097/MAO.0000000000002710. Accessed: Jan. 15, 2024. [Online]. Available: https://journals.lww.com/otology-neurotology/abstract/2020/09000/predicting_postoperative_cochlear_implant.16.aspx.
- [149] F. Milella, L. Famiglini, G. Banfi, and F. Cabitza, "Application of Machine Learning to Improve Appropriateness of Treatment in an Orthopaedic Setting of Personalized Medicine," eng, *Journal of Personalized Medicine*, vol. 12, no. 10, p. 1706, Oct. 2022, ISSN: 2075-4426. DOI: 10.3390/jpm12101706.
- [150] E. M. Polce et al., "Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty," English, *Journal of Shoulder and Elbow Surgery*, vol. 30, no. 6, e290–e299, Jun. 2021, Publisher: Elsevier, ISSN: 1058-2746, 1532-6500. DOI: 10.1016/j.jse.2020.09.007. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.jshoulderelbow.org/article/S1058-2746\(20\)30769-2/fulltext](https://www.jshoulderelbow.org/article/S1058-2746(20)30769-2/fulltext).
- [151] P. Thanathamathsee, "Boosting with feature selection technique for screening and predicting adolescents depression," in *2014 Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP)*, May 2014, pp. 23–27. DOI: 10.1109/DICTAP.2014.6821650. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/6821650>.
- [152] J. A. A. Mendoza et al., "A Machine Learning Approach in Evaluating Symptom Screening in Predicting COVID-19," in *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, Feb. 2022, pp. 188–193. DOI: 10.1109/ICAIIIC54071.2022.9722659. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9722659>.
- [153] Z. Xu, Y. Wang, and B. Jiang, "Research on Prisoner Psychological Symptoms Quick Screening Model Based on Ensemble Learning," in *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*, ser. ISAIMS '22, New York, NY, USA: Association for Computing Machinery, Dec. 2022, pp. 563–567, ISBN: 978-1-4503-9844-2. DOI: 10.1145/3570773.3570870. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1145/3570773.3570870>.
- [154] M. R. Fu et al., "Machine learning for detection of lymphedema among breast cancer survivors," en, *mHealth*, vol. 4, no. 5, May 2018, Publisher: AME publishing company, ISSN: 2306-9740. DOI: 10.21037/mhealth.2018.04.02.

-
- Accessed: Feb. 28, 2023. [Online]. Available: <https://mhealth.amegroups.com/article/view/19663>.
- [155] X. Pan et al., “Dosimetric predictors of patient-reported toxicity after prostate stereotactic body radiotherapy: Analysis of full range of the dose–volume histogram using ensemble machine learning,” English, *Radiotherapy and Oncology*, vol. 148, pp. 181–188, Jul. 2020, Publisher: Elsevier, ISSN: 0167-8140, 1879-0887. DOI: 10.1016/j.radonc.2020.04.013. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.thegreenjournal.com/article/S0167-8140\(20\)30194-8/fulltext](https://www.thegreenjournal.com/article/S0167-8140(20)30194-8/fulltext).
 - [156] N. Agochukwu-Mmonu et al., “Development and Validation of Dynamic Multivariate Prediction Models of Sexual Function Recovery in Patients with Prostate Cancer Undergoing Radical Prostatectomy: Results from the MUSIC Statewide Collaborative,” *European Urology Open Science*, vol. 40, pp. 1–8, Jun. 2022, ISSN: 2666-1683. DOI: 10.1016/j.euros.2022.03.009. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666168322000623>.
 - [157] A. M. Chekroud et al., “Cross-trial prediction of treatment outcome in depression: A machine learning approach,” English, *The Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, Mar. 2016, Publisher: Elsevier, ISSN: 2215-0366, 2215-0374. DOI: 10.1016/S2215-0366(15)00471-X. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366\(15\)00471-X/fulltext](https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(15)00471-X/fulltext).
 - [158] W. Oude Nijeweme-d’Hollosy, L. van Velsen, M. Poel, C. G. M. Groothuis-Oudshoorn, R. Soer, and H. Hermens, “Evaluation of three machine learning models for self-referral decision support on low back pain in primary care,” *International Journal of Medical Informatics*, vol. 110, pp. 31–41, Feb. 2018, ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2017.11.010. Accessed: Dec. 31, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386505617304240>.
 - [159] A. Goldstein and S. Cohen, “Self-report symptom-based endometriosis prediction using machine learning,” en, *Scientific Reports*, vol. 13, no. 1, p. 5499, Apr. 2023, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-023-32761-8. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.nature.com/articles/s41598-023-32761-8>.
 - [160] S. Iivanainen, J. Ekström, H. Virtanen, V. V. Kataja, and J. P. Koivunen, “Predicting Objective Response Rate (ORR) in Immune Checkpoint Inhibitor (ICI) Therapies with Machine Learning (ML) by Combining Clinical and Patient-Reported Data,” en, *Applied Sciences*, vol. 12, no. 3, p. 1563, Jan. 2022, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app12031563. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.mdpi.com/2076-3417/12/3/1563>.
 - [161] Y. Luo, “Evaluating the state of the art in missing data imputation for clinical data,” *Briefings in Bioinformatics*, vol. 23, no. 1, bbab489, Jan. 2022, ISSN: 1477-4054. DOI: 10.1093/bib/bbab489. Accessed: Jul. 5, 2023. [Online]. Available: <https://doi.org/10.1093/bib/bbab489>.
-

-
- [162] H. S. Haeberle et al., “Predicting the Risk of Subsequent Hip Surgery Before Primary Hip Arthroscopy for Femoroacetabular Impingement Syndrome: A Machine Learning Analysis of Preoperative Risk Factors in Hip Preservation,” eng, *The American Journal of Sports Medicine*, vol. 49, no. 10, pp. 2668–2676, Aug. 2021, ISSN: 1552-3365. DOI: 10.1177/03635465211024964.
- [163] R. Sun, C. Tomkins-Lane, A. Muaremi, A. Kuwabara, and M. Smuck, “Physical activity thresholds for predicting longitudinal gait decline in adults with knee osteoarthritis,” eng, *Osteoarthritis and Cartilage*, vol. 29, no. 7, pp. 965–972, Jul. 2021, ISSN: 1522-9653. DOI: 10.1016/j.joca.2021.04.002.
- [164] K. Schultebrasucks et al., “Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: A machine-learning approach for analyzing multivariate predictors,” en, *Molecular Psychiatry*, vol. 26, no. 9, pp. 5011–5022, Sep. 2021, Number: 9 Publisher: Nature Publishing Group, ISSN: 1476-5578. DOI: 10.1038/s41380-020-0789-2. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.nature.com/articles/s41380-020-0789-2>.
- [165] M. H. Sandham, E. A. Hedgecock, R. J. Siegert, A. Narayanan, M. B. Hocaoglu, and I. J. Higginson, “Intelligent Palliative Care Based on Patient-Reported Outcome Measures,” English, *Journal of Pain and Symptom Management*, vol. 63, no. 5, pp. 747–757, May 2022, Publisher: Elsevier, ISSN: 0885-3924, 1873-6513. DOI: 10.1016/j.jpainsymman.2021.11.008. Accessed: Jan. 15, 2024. [Online]. Available: [https://www.jpmsjournal.com/article/S0885-3924\(21\)00641-2/fulltext](https://www.jpmsjournal.com/article/S0885-3924(21)00641-2/fulltext).
- [166] H. Pappot et al., “Machine learning applied in patient-reported outcome research—exploring symptoms in adjuvant treatment of breast cancer,” en, *Breast Cancer*, vol. 31, no. 1, pp. 148–153, Jan. 2024, ISSN: 1880-4233. DOI: 10.1007/s12282-023-01515-9. Accessed: Jan. 15, 2024. [Online]. Available: <https://doi.org/10.1007/s12282-023-01515-9>.
- [167] Y.-C. Hsu, J.-D. Wang, P.-H. Huang, Y.-W. Chien, C.-J. Chiu, and C.-Y. Lin, “Integrating domain knowledge with machine learning to detect obstructive sleep apnea: Snore as a significant bio-feature,” en, *Journal of Sleep Research*, vol. 31, no. 2, e13487, 2022, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.13487>, ISSN: 1365-2869. DOI: 10.1111/jsr.13487. Accessed: Jan. 15, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.13487>.
- [168] I. Miranda, G. Cardoso, M. Pahar, G. Oliveira, and T. Niesler, “Machine Learning Prediction of Hospitalization due to COVID-19 based on Self-Reported Symptoms: A Study for Brazil,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, ISSN: 2641-3604, Jul. 2021, pp. 1–5. DOI: 10.1109/BHI50953.2021.9508548. Accessed: Jan. 15, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9508548>.

-
- [169] M. J. Long, E. Papi, L. D. Duffell, and A. H. McGregor, "Predicting knee osteoarthritis risk in injured populations," *Clinical Biomechanics*, vol. 47, pp. 87–95, Aug. 2017, ISSN: 0268-0033. DOI: 10.1016/j.clinbiomech.2017.06.001. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0268003317301286>.
- [170] W. Xuyi, H. Seow, and R. Sutradhar, "Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer," *Cancer Medicine*, vol. 10, no. 3, pp. 989–998, Dec. 2020, ISSN: 2045-7634. DOI: 10.1002/cam4.3685. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7897969/>.
- [171] M. N. Shahzad, M. Suleman, M. A. Ahmed, A. Riaz, and K. Fatima, "Identifying the Symptom Severity in Obsessive-Compulsive Disorder for Classification and Prediction: An Artificial Neural Network Approach," *Behavioural Neurology*, vol. 2020, p. 2678718, Jun. 2020, ISSN: 0953-4180. DOI: 10.1155/2020/2678718. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327615/>.
- [172] A. Bugajski, A. Lengerich, R. Koerner, and L. Szalacha, "Utilizing an Artificial Neural Network to Predict Self-Management in Patients With Chronic Obstructive Pulmonary Disease: An Exploratory Analysis," en, *Journal of Nursing Scholarship*, vol. 53, no. 1, pp. 16–24, 2021, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnu.12618>, ISSN: 1547-5069. DOI: 10.1111/jnu.12618. Accessed: Jan. 15, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jnu.12618>.
- [173] L. van der Stap et al., "The feasibility of a Bayesian network model to assess the probability of simultaneous symptoms in patients with advanced cancer," en, *Scientific Reports*, vol. 12, no. 1, p. 22295, Dec. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-022-26342-4. Accessed: Jan. 15, 2024. [Online]. Available: <https://www.nature.com/articles/s41598-022-26342-4>.
- [174] T. Strating, L. Shafiee Hanjani, I. Tornvall, R. Hubbard, and I. A. Scott, "Navigating the machine learning pipeline: A scoping review of inpatient delirium prediction models," *BMJ Health & Care Informatics*, vol. 30, no. 1, e100767, Jul. 2023, ISSN: 2632-1009. DOI: 10.1136/bmjhci-2023-100767. Accessed: Jan. 3, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10335592/>.
- [175] S. Cruz Rivera et al., "Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies," en, *The Lancet Digital Health*, vol. 5, no. 3, e168–e173, Mar. 2023, ISSN: 25897500. DOI: 10.1016/S2589-7500(22)00252-7. Accessed: Dec. 12, 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2589750022002527>.
- [176] R. R. Whitebird et al., "What Do Orthopaedists Believe is Needed for Incorporating Patient-reported Outcome Measures into Clinical Care? A Qualitative Study," *Clinical Orthopaedics and Related Research*, vol. 480, no. 4, pp. 680–687, Apr. 2022, ISSN: 0009-921X. DOI: 10.1097/CORR.0000000000002059. Accessed: Dec. 12, 2023. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8923577/>.
-

-
- [177] M. R. Davies et al., “Comparison of symptom-based versus self-reported diagnostic measures of anxiety and depression disorders in the GLAD and COPING cohorts,” *Journal of Anxiety Disorders*, vol. 85, p. 102491, Jan. 2022, ISSN: 0887-6185. DOI: 10.1016/j.janxdis.2021.102491. Accessed: Jan. 4, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0887618521001389>.
- [178] S. Schmeelk et al., “Monitoring Symptoms of COVID-19: Review of Mobile Apps,” EN, *JMIR mHealth and uHealth*, vol. 10, no. 6, e36065, Jun. 2022, Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada. DOI: 10.2196/36065. Accessed: Jan. 4, 2024. [Online]. Available: <https://mhealth.jmir.org/2022/6/e36065>.
- [179] M. Sharma, C. Savage, M. Nair, I. Larsson, P. Svedberg, and J. M. Nygren, “Artificial Intelligence Applications in Health Care Practice: Scoping Review,” en, *Journal of Medical Internet Research*, vol. 24, no. 10, e40238, Oct. 2022, ISSN: 1438-8871. DOI: 10.2196/40238. Accessed: Jan. 5, 2024. [Online]. Available: <https://www.jmir.org/2022/10/e40238>.
- [180] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons, “Review: A gentle introduction to imputation of missing values,” *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1087–1091, Oct. 2006, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2006.01.014. Accessed: Jan. 4, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895435606001971>.
- [181] M. Pe et al., “Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer Clinical Trials-Innovative Medicines Initiative (SISAQOL-IMI): Stakeholder views, objectives, and procedures,” eng, *The Lancet. Oncology*, vol. 24, no. 6, e270–e283, Jun. 2023, ISSN: 1474-5488. DOI: 10.1016/S1470-2045(23)00157-2.
- [182] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, “K-Nearest Neighbor (K-NN) based Missing Data Imputation,” in *2019 5th International Conference on Science in Information Technology (ICSITech)*, Oct. 2019, pp. 83–88. DOI: 10.1109/ICSITech46713.2019.8987530. Accessed: Jan. 4, 2024. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8987530>.
- [183] Y. Liu, Y. Li, and D. Xie, “Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks,” *Journal of Statistical Computation and Simulation*, vol. 94, no. 1, pp. 183–203, Jan. 2024, Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/00949655.2023.2238235>, ISSN: 0094-9655. DOI: 10.1080/00949655.2023.2238235. Accessed: Jan. 4, 2024. [Online]. Available: <https://doi.org/10.1080/00949655.2023.2238235>.
- [184] D. Jha et al., *Ensuring Trustworthy Medical Artificial Intelligence through Ethical and Philosophical Principles*, arXiv:2304.11530 [cs], Sep. 2023. Accessed: Jan. 4, 2024. [Online]. Available: <http://arxiv.org/abs/2304.11530>.
-

-
- [185] D.-E.-M. Nisar, R. Amin, N.-U.-H. Shah, M. A. A. Ghamdi, S. H. Almotiri, and M. Alruily, "Healthcare Techniques Through Deep Learning: Issues, Challenges and Opportunities," en, *IEEE Access*, vol. 9, pp. 98 523–98 541, 2021, ISSN: 2169-3536. DOI: 10 . 1109 / ACCESS . 2021 . 3095312. Accessed: Jan. 4, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/9476037/>.
- [186] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction Using Deep Learning Methods in Healthcare," en, *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–29, Mar. 2023, ISSN: 2158-656X, 2158-6578. DOI: 10.1145/3531326. Accessed: Jun. 27, 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3531326>.
- [187] I. R. Ward, L. Wang, J. lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable Artificial Intelligence for Pharmacovigilance: What Features Are Important When Predicting Adverse Outcomes?" *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106 415, Nov. 2021, arXiv:2112.13210 [cs, q-bio], ISSN: 01692607. DOI: 10 . 1016 / j . cmpb . 2021 . 106415. Accessed: Jan. 4, 2024. [Online]. Available: <http://arxiv.org/abs/2112.13210>.
- [188] C. Sidey-Gibbons et al., "Development of Machine Learning Algorithms for the Prediction of Financial Toxicity in Localized Breast Cancer Following Surgical Treatment," en, *JCO Clinical Cancer Informatics*, no. 5, pp. 338–347, Dec. 2021, ISSN: 2473-4276. DOI: 10 . 1200 / CCI . 20 . 00088. Accessed: Aug. 29, 2022. [Online]. Available: <https://ascopubs.org/doi/10.1200/CCI.20.00088>.
- [189] B. S. Kelly et al., "The ethical matrix as a method for involving people living with disease and the wider public (PPI) in near-term artificial intelligence research," *Radiography, Patient Voice and the Patient Experience*, vol. 29, S103–S111, May 2023, ISSN: 1078-8174. DOI: 10.1016/j.radi.2023.03.009. Accessed: Jan. 5, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1078817423000767>.
- [190] Y. Chen, A. A. Hosin, M. J. George, F. W. Asselbergs, and A. D. Shah, "Digital technology and patient and public involvement (PPI) in routine care and clinical research—A pilot study," en, *PLOS ONE*, vol. 18, no. 2, e0278260, Feb. 2023, ISSN: 1932-6203. DOI: 10 . 1371 / journal . pone . 0278260. Accessed: Jan. 5, 2024. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0278260>.
- [191] M. Yu, M. Harrison, and N. Bansback, "Can prediction models for hospital readmission be improved by incorporating patient-reported outcome measures? A systematic review and narrative synthesis," en, *Quality of Life Research*, vol. 33, no. 7, pp. 1767–1779, Jul. 2024, ISSN: 1573-2649. DOI: 10 . 1007 / s11136 - 024 - 03638 - 8. Accessed: Aug. 21, 2025. [Online]. Available: <https://doi.org/10.1007/s11136-024-03638-8>.
- [192] L. S. Lundgren, N. Willems, R. C. Marchand, C. Batailler, and S. Lustig, "Surgical factors play a critical role in predicting functional outcomes using machine learning in robotic-assisted total knee arthroplasty," en, *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 32, no. 12, pp. 3198–3209, 2024, ISSN: 1433-7347. DOI: 10 . 1002 / ksa . 12302. Accessed: Aug. 21, 2025.

-
- [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ksa.12302>.
- [193] J. Park et al., "Machine Learning-Based Predictive Models for 90-Day Readmission of Total Joint Arthroplasty Using Comprehensive Electronic Health Records and Patient-Reported Outcome Measures," *Arthroplasty Today*, vol. 25, p. 101308, Feb. 2024, ISSN: 2352-3441. DOI: 10.1016/j.artd.2023.101308. Accessed: Aug. 21, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352344123002133>.
- [194] Y. Mahmoud, K. Horvath, and Y. Zhou, "Deep Learning for Predicting Rehabilitation Success: Advancing Clinical and Patient-Reported Outcome Modeling," en, *Electronics*, vol. 14, no. 6, p. 1082, Jan. 2025, Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics14061082. Accessed: Aug. 21, 2025. [Online]. Available: <https://www.mdpi.com/2079-9292/14/6/1082>.
- [195] L. Beenhakker et al., "Development of machine learning models to predict cancer-related fatigue in Dutch breast cancer survivors up to 15 years after diagnosis," en, *Journal of Cancer Survivorship*, vol. 19, no. 2, pp. 580–593, Apr. 2025, ISSN: 1932-2267. DOI: 10.1007/s11764-023-01491-1. Accessed: Aug. 21, 2025. [Online]. Available: <https://doi.org/10.1007/s11764-023-01491-1>.
- [196] D. T. Eton, D. L. Fairclough, D. Cella, S. E. Yount, P. Bonomi, and D. H. Johnson, "Early Change in Patient-Reported Health During Lung Cancer Chemotherapy Predicts Clinical Outcomes Beyond Those Predicted by Baseline Report: Results From Eastern Cooperative Oncology Group Study 5592," *Journal of Clinical Oncology*, vol. 21, no. 8, pp. 1536–1543, Apr. 2003, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/JCO.2003.07.128. Accessed: Sep. 2, 2022. [Online]. Available: <https://ascopubs.org/doi/abs/10.1200/JCO.2003.07.128>.
- [197] J. A. McKnight, "Principles of chemotherapy," *Clinical Techniques in Small Animal Practice*, Medical Oncology, vol. 18, no. 2, pp. 67–72, May 2003, ISSN: 1096-2867. DOI: 10.1053/svms.2003.36617. Accessed: Sep. 4, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1096286703800026>.
- [198] P. Di Nardo et al., "Chemotherapy in patients with early breast cancer: Clinical overview and management of long-term side effects," *Expert Opinion on Drug Safety*, vol. 21, no. 11, pp. 1341–1355, Nov. 2022, Publisher: Taylor & Francis, eprint: <https://doi.org/10.1080/14740338.2022.2151584>, ISSN: 1474-0338. DOI: 10.1080/14740338.2022.2151584. Accessed: Jan. 9, 2025. [Online]. Available: <https://doi.org/10.1080/14740338.2022.2151584>.
- [199] I. Ingrand et al., "Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey," en, *British Journal of Clinical Pharmacology*, vol. 86, no. 4, pp. 711–722, 2020, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bcp.14159>, ISSN: 1365-2125. DOI: 10.1111/bcp.14159. Accessed: Jan. 9, 2025. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bcp.14159>.
-

-
- [200] O. Carrasquillo, "Health Care Utilization," en, in *Encyclopedia of Behavioral Medicine*, M. D. Gellman and J. R. Turner, Eds., New York, NY: Springer, 2013, pp. 909–910, ISBN: 978-1-4419-1005-9. DOI: 10.1007/978-1-4419-1005-9_885. Accessed: Apr. 10, 2025. [Online]. Available: https://doi.org/10.1007/978-1-4419-1005-9_885.
- [201] K. N. Lohr and B. J. Zebrack, "Using patient-reported outcomes in clinical practice: Challenges and opportunities," en, *Quality of Life Research*, vol. 18, no. 1, pp. 99–107, Feb. 2009, ISSN: 1573-2649. DOI: 10.1007/s11136-008-9413-7. Accessed: May 9, 2025. [Online]. Available: <https://doi.org/10.1007/s11136-008-9413-7>.
- [202] P. Dolan, "Modeling valuations for EuroQol health states," eng, *Medical Care*, vol. 35, no. 11, pp. 1095–1108, Nov. 1997, ISSN: 0025-7079. DOI: 10.1097/00005650-199711000-00002.
- [203] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," en, Apr. 2024, Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting. DOI: 10.1136/bmj-2023-078378. Accessed: Sep. 9, 2025. [Online]. Available: <https://www.bmj.com/content/385/bmj-2023-078378>.
- [204] T. Horsley, "Tips for Improving the Writing and Reporting Quality of Systematic, Scoping, and Narrative Reviews," en-US, *Journal of Continuing Education in the Health Professions*, vol. 39, no. 1, p. 54, 2019, ISSN: 0894-1912. DOI: 10.1097/CEH.0000000000000241. Accessed: Aug. 20, 2025. [Online]. Available: https://journals.lww.com/jcehp/fulltext/2019/03910/Tips_for_Improving_the_Writing_and_Reporting.10.aspx.
- [205] Z. Munn, M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 143, Nov. 2018, ISSN: 1471-2288. DOI: 10.1186/s12874-018-0611-x. Accessed: Aug. 12, 2025. [Online]. Available: <https://doi.org/10.1186/s12874-018-0611-x>.
- [206] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.11.011. Accessed: Jan. 9, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253521002360>.
- [207] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," *Expert Systems with Applications*, vol. 182, p. 115222, Nov. 2021, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2021.115222. Accessed: Apr. 10, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417421006540>.
- [208] Y. Wang et al., "Improving Prediction of Late Symptoms using LSTM and Patient-reported Outcomes for Head and Neck Cancer Patients," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, ISSN: 2575-2634, Jun. 2023, pp. 292–300. DOI: 10.1109/ICHI57859.2023.00047.

-
- Accessed: Jan. 9, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10337216>.
- [209] P. M. Livingston, M. Craike, and J. Considine, “Unplanned presentations to emergency departments due to chemotherapy induced complications: Opportunities for improving service delivery,” en, *Australasian Emergency Nursing Journal*, vol. 14, no. 2, pp. 62–68, May 2011, ISSN: 1574-6267. DOI: 10.1016/j.aenj.2011.03.005. Accessed: Aug. 28, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574626711000413>.
 - [210] M. T. Amjad, A. Chidharla, and A. Kasi, “Cancer Chemotherapy,” eng, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2022. Accessed: Sep. 2, 2022. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK564367/>.
 - [211] P. H. Dufton, M. F. Gerdtz, R. Jarden, and M. Krishnasamy, “Methodological approaches to measuring the incidence of unplanned emergency department presentations by cancer patients receiving systemic anti-cancer therapy: A systematic review,” en, *BMC Medical Research Methodology*, vol. 22, no. 1, p. 75, Dec. 2022, ISSN: 1471-2288. DOI: 10.1186/s12874-022-01555-3. Accessed: Aug. 23, 2022. [Online]. Available: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01555-3>.
 - [212] J. C. Hong, D. Niedzwiecki, M. Palta, and J. D. Tenenbaum, “Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm,” en, *JCO Clinical Cancer Informatics*, no. 2, pp. 1–11, Dec. 2018, ISSN: 2473-4276. DOI: 10.1200/CCI.18.00037.
 - [213] J. Feliu et al., “Prediction of Chemotoxicity, Unplanned Hospitalizations and Early Death in Older Patients with Colorectal Cancer Treated with Chemotherapy,” en, *Cancers*, vol. 14, no. 1, p. 127, Jan. 2022, Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-6694. DOI: 10.3390/cancers14010127. Accessed: Jan. 9, 2025. [Online]. Available: <https://www.mdpi.com/2072-6694/14/1/127>.
 - [214] M. Pisu and M. Y. Martin, “Financial toxicity: A common problem affecting patient care and health,” *Nature reviews. Disease primers*, vol. 8, no. 1, p. 7, Feb. 2022, ISSN: 2056-676X. DOI: 10.1038/s41572-022-00341-1. Accessed: Jun. 18, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9731797/>.
 - [215] S. Haneuse, D. Arterburn, and M. J. Daniels, “Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task,” *JAMA Network Open*, vol. 4, no. 2, e210184, Feb. 2021, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.0184. Accessed: Jul. 5, 2023. [Online]. Available: <https://doi.org/10.1001/jamanetworkopen.2021.0184>.
 - [216] D. A. Bennett, “How can I deal with missing data in my study?” en, *Australian and New Zealand Journal of Public Health*, vol. 25, no. 5, pp. 464–469, 2001, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-842X.2001.tb00294.x>, ISSN: 1753-6405. DOI: 10.1111/j.1467-842X.2001.tb00294.x. Accessed: Jun. 5, 2023. [Online]. Available:
-

<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2001.tb00294.x>.

- [217] T. L. Wiemken and R. R. Kelley, “Machine Learning in Epidemiology and Health Outcomes Research,” en, *Annual Review of Public Health*, vol. 41, no. 1, pp. 21–36, Apr. 2020, ISSN: 0163-7525, 1545-2093. DOI: 10.1146/annurev-publhealth-040119-094437. Accessed: Jun. 5, 2023. [Online]. Available: <https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040119-094437>.
- [218] S. W. a. C. S. a. M. Price, *Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach — EndNote Click*, en. Accessed: Jun. 6, 2023. [Online]. Available: <https://click.endnote.com/viewer?doi=10.2196%2F13946&token=WzQwMTgyOTYsIjEwLjIxOTYvMTM5NDYiXQ.MLoVAYxp7bQ2bc6AZqm2b0P41VA>.
- [219] L. Goh, Q. Song, and N. Kasabov, “A Novel Feature Selection Method to Improve Classification of Gene Expression Data,” en,
- [220] M. Savić et al., “Analysis of Machine Learning Models Predicting Quality of Life for Cancer Patients,” en, in *Proceedings of the 13th International Conference on Management of Digital EcoSystems*, Virtual Event Tunisia: ACM, Nov. 2021, pp. 35–42, ISBN: 978-1-4503-8314-1. DOI: 10.1145/3444757.3485103. Accessed: Jun. 9, 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3444757.3485103>.
- [221] W. L. Cava, C. Bauer, J. H. Moore, and S. A. Pendergrass, “Interpretation of machine learning predictions for patient outcomes in electronic health records,” *AMIA Annual Symposium Proceedings*, vol. 2019, pp. 572–581, Mar. 2020, ISSN: 1942-597X. Accessed: Jun. 20, 2023. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153071/>.
- [222] B. Langenberger, A. Thoma, and V. Vogt, “Can minimal clinically important differences in patient reported outcome measures be predicted by machine learning in patients with total knee or hip arthroplasty? A systematic review,” *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 18, Jan. 2022, ISSN: 1472-6947. DOI: 10.1186/s12911-022-01751-7. Accessed: Jun. 9, 2023. [Online]. Available: <https://doi.org/10.1186/s12911-022-01751-7>.
- [223] D. J. Peterson, N. P. Ostberg, D. W. Blayney, J. D. Brooks, and T. Hernandez-Boussard, “Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions,” en, *JCO Clinical Cancer Informatics*, no. 5, pp. 1106–1126, Dec. 2021, ISSN: 2473-4276. DOI: 10.1200/CCI.21.00116. Accessed: Feb. 28, 2023. [Online]. Available: <https://ascopubs.org/doi/10.1200/CCI.21.00116>.
- [224] S. H. Kim, M.-W. Jo, H.-J. Kim, and J.-H. Ahn, “Mapping EORTC QLQ-C30 onto EQ-5D for the assessment of cancer patients,” *Health and Quality of Life Outcomes*, vol. 10, no. 1, p. 151, Dec. 2012, ISSN: 1477-7525. DOI: 10.1186/1477-7525-10-151. Accessed: Jun. 12, 2023. [Online]. Available: <https://doi.org/10.1186/1477-7525-10-151>.

-
- [225] V. Azimi and M. A. Zaydman, “Optimizing Equity: Working towards Fair Machine Learning Algorithms in Laboratory Medicine,” *The Journal of Applied Laboratory Medicine*, vol. 8, no. 1, pp. 113–128, Jan. 2023, ISSN: 2576-9456. DOI: 10.1093/jalm/jfac085. Accessed: Jun. 13, 2023. [Online]. Available: <https://doi.org/10.1093/jalm/jfac085>.
- [226] R. Shafique et al., “Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning,” en, *Cancers*, vol. 15, no. 3, p. 681, Jan. 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15030681. Accessed: Jun. 13, 2023. [Online]. Available: <https://www.mdpi.com/2072-6694/15/3/681>.
- [227] T. Davenport and R. Kalakota, “The potential for artificial intelligence in healthcare,” *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, Jun. 2019, ISSN: 2514-6645. DOI: 10.7861/futurehosp.6-2-94. Accessed: Jun. 13, 2023. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/>.
- [228] L. Blonde, K. Khunti, S. B. Harris, C. Meizinger, and N. S. Skolnik, “Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician,” en, *Advances in Therapy*, vol. 35, no. 11, pp. 1763–1774, Nov. 2018, ISSN: 1865-8652. DOI: 10.1007/s12325-018-0805-y. Accessed: Aug. 30, 2022. [Online]. Available: <https://doi.org/10.1007/s12325-018-0805-y>.
- [229] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” en, *PLOS ONE*, vol. 14, no. 11, e0224365, Nov. 2019, Publisher: Public Library of Science, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0224365. Accessed: Aug. 30, 2022. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365>.
- [230] D. Krepper, M. Cesari, N. J. Hubel, P. Zelger, and M. J. Sztankay, “Machine learning models including patient-reported outcome data in oncology: A systematic literature review and analysis of their reporting quality,” *Journal of Patient-Reported Outcomes*, vol. 8, no. 1, p. 126, Nov. 2024, ISSN: 2509-8020. DOI: 10.1186/s41687-024-00808-7. Accessed: Jan. 9, 2025. [Online]. Available: <https://doi.org/10.1186/s41687-024-00808-7>.
- [231] W. Linden, A. Vodermaier, R. MacKenzie, and D. Greig, “Anxiety and depression after cancer diagnosis: Prevalence rates by cancer type, gender, and age,” en, *Journal of Affective Disorders*, vol. 141, no. 2, pp. 343–351, Dec. 2012, ISSN: 0165-0327. DOI: 10.1016/j.jad.2012.03.025. Accessed: Jun. 28, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0165032712002212>.
- [232] M. Kowal, F. Douglas, D. Jayne, and D. Meads, “Patient choice in colorectal cancer treatment – A systematic review and narrative synthesis of attribute-based stated preference studies,” en, *Colorectal Disease*, vol. 24, no. 11, pp. 1295–1307, 2022, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/codi.16242>, ISSN: 1463-1318. DOI: 10.1111/codi.16242. Accessed: Jun. 28, 2023. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/codi.16242>.
-

-
- [233] A. Shrestha, C. Martin, M. Burton, S. Walters, K. Collins, and L. Wyld, “Quality of life versus length of life considerations in cancer patients: A systematic literature review,” en, *Psycho-Oncology*, vol. 28, no. 7, pp. 1367–1380, 2019, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/pon.5054>, ISSN: 1099-1611. DOI: 10.1002/pon.5054. Accessed: Jun. 28, 2023. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/pon.5054>.
- [234] R. Matsuyama, S. Reddy, and T. J. Smith, “Why Do Patients Choose Chemotherapy Near the End of Life? A Review of the Perspective of Those Facing Death From Cancer,” en, *Journal of Clinical Oncology*, vol. 24, no. 21, pp. 3490–3496, Jul. 2006, ISSN: 0732-183X, 1527-7755. DOI: 10.1200/JCO.2005.03.6236. Accessed: Jun. 27, 2023. [Online]. Available: <https://ascopubs.org/doi/10.1200/JCO.2005.03.6236>.
- [235] S. Chae, W. N. Street, N. Ramaraju, and S. Gilbertson-White, “Prediction of Cancer Symptom Trajectory Using Longitudinal Electronic Health Record Data and Long Short-Term Memory Neural Network,” *JCO Clinical Cancer Informatics*, no. 8, e2300039, Mar. 2024, Publisher: Wolters Kluwer. DOI: 10.1200/CCI.23.00039. Accessed: Jan. 9, 2025. [Online]. Available: <https://ascopubs.org/doi/abs/10.1200/CCI.23.00039>.
- [236] K. Zhou, M. Bellanger, S. Le Lann, M. Robert, J.-S. Frenel, and M. Campone, “The predictive value of patient-reported outcomes on the impact of breast cancer treatment-related quality of life,” eng, *Frontiers in Oncology*, vol. 12, p. 925 534, 2022, ISSN: 2234-943X. DOI: 10.3389/fonc.2022.925534.
- [237] M. T. King et al., “Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G,” *Journal of Clinical Epidemiology*, vol. 63, no. 3, pp. 270–281, Mar. 2010, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2009.05.001. Accessed: Mar. 9, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895435609001498>.
- [238] J. Z. Musoro et al., “Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer,” *JNCI Cancer Spectrum*, vol. 3, no. 3, pkz037, Jun. 2019, ISSN: 2515-5091. DOI: 10.1093/jncics/pkz037. Accessed: Mar. 9, 2024. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7050000/>.
- [239] T. Aljrees, “Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning,” en, *PLOS ONE*, vol. 19, no. 1, e0295632, Jan. 2024, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0295632. Accessed: Mar. 27, 2024. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0295632>.
- [240] S. Han, T. J. Sohn, B. P. Ng, and C. Park, “Predicting unplanned readmission due to cardiovascular disease in hospitalized patients with cancer: A machine learning approach,” en, *Scientific Reports*, vol. 13, no. 1, p. 13 491, Aug. 2023, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-023-40552-4. Accessed: Jan. 10, 2025. [Online]. Available: <https://www.nature.com/articles/s41598-023-40552-4>.

-
- [241] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, Jun. 2019, ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2019.02.004. Accessed: Apr. 4, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895435618310813>.
- [242] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” en, *BMC Bioinformatics*, vol. 7, no. 1, p. 91, Feb. 2006, ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-91. Accessed: Jan. 10, 2025. [Online]. Available: <https://doi.org/10.1186/1471-2105-7-91>.
- [243] R. Muthukrishnan and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Oct. 2016, pp. 18–20. DOI: 10.1109/ICACA.2016.7887916.
- [244] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” en, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. Accessed: Jun. 30, 2025. [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10302>.
- [245] N. S. Rahmi, N. W. S. Wardhani, M. B. Mitakda, R. S. Fauztina, and I. Salsabila, “SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data : (Case Study of Early Detection of Cervical Cancer in Indonesia),” in *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Nov. 2022, pp. 1–6. DOI: 10.1109/ICITDA55840.2022.9971421. Accessed: Jun. 30, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/9971421>.
- [246] S. Kapoor and A. Narayanan, *Leakage and the Reproducibility Crisis in ML-based Science*, arXiv:2207.07048 [cs], Jul. 2022. DOI: 10.48550/arXiv.2207.07048. Accessed: Jan. 9, 2025. [Online]. Available: <http://arxiv.org/abs/2207.07048>.
- [247] P. Thölke et al., “Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data,” *NeuroImage*, vol. 277, p. 120 253, Aug. 2023, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2023.120253. Accessed: Jul. 1, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811923004044>.
- [248] R. A. Fisher, “Statistical Methods for Research Workers,” in *Breakthroughs in Statistics*, S. Kotz and N. L. Johnson, Eds., Series Title: Springer Series in Statistics, New York, NY: Springer New York, 1992, pp. 66–70, ISBN: 978-0-387-94039-7 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_6. Accessed: Aug. 15, 2025. [Online]. Available: http://link.springer.com/10.1007/978-1-4612-4380-9_6.

-
- [249] W. K. Wang et al., “A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications,” en, *Sensors*, vol. 22, no. 20, p. 8016, Jan. 2022, Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1424-8220. DOI: 10.3390/s22208016. Accessed: Aug. 15, 2025. [Online]. Available: <https://www.mdpi.com/1424-8220/22/20/8016>.
- [250] A. Wilson and M. R. Anwar, “The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing,” en, *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 97–107, Nov. 2024, ISSN: 2963-1939. DOI: 10.33050/italic.v3i1.656. Accessed: Aug. 15, 2025. [Online]. Available: <https://journal.pandawan.id/italic/article/view/656>.
- [251] L. Tang et al., “Implementing Symptom Management Follow-up Using an Electronic Patient-Reported Outcome Platform in Outpatients With Advanced Cancer: Longitudinal Single-Center Prospective Study,” EN, *JMIR Formative Research*, vol. 6, no. 5, e21458, May 2022, Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research Publisher: JMIR Publications Inc., Toronto, Canada. DOI: 10.2196/21458. Accessed: Jul. 3, 2025. [Online]. Available: <https://formative.jmir.org/2022/5/e21458>.
- [252] G. A. Brooks, A. J. Kansagra, S. R. Rao, J. I. Weitzman, E. A. Linden, and J. O. Jacobson, “A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy,” *JAMA oncology*, vol. 1, no. 4, pp. 441–447, Jul. 2015, ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.0828. Accessed: Sep. 10, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576342/>.
- [253] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, “A review of irregular time series data handling with gated recurrent neural networks,” *Neurocomputing*, vol. 441, pp. 161–178, Jun. 2021, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.02.046. Accessed: Jan. 9, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231221003003>.
- [254] D. Li, P. Lyons, J. Klaus, B. Gage, M. Kollef, and C. Lu, “Integrating Static and Time-Series Data in Deep Recurrent Models for Oncology Early Warning Systems,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM ’21, New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 913–936, ISBN: 978-1-4503-8446-9. DOI: 10.1145/3459637.3482441. Accessed: Jan. 9, 2025. [Online]. Available: <https://dl.acm.org/doi/10.1145/3459637.3482441>.
- [255] C. Lin et al., “Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, ISSN: 2575-2634, Jun. 2018, pp. 219–228. DOI: 10.1109/ICHI.2018.00032. Accessed: Jan. 9, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/8419365>.

-
- [256] T. Zidaru, E. M. Morrow, and R. Stockley, “Ensuring patient and public involvement in the transition to AI-assisted mental health care: A systematic scoping review and agenda for design justice,” en, *Health Expectations*, vol. 24, no. 4, pp. 1072–1124, 2021, ISSN: 1369-7625. DOI: 10.1111/hex.13299. Accessed: Jul. 6, 2023.
- [257] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, and N. D. B. Bruce, “Position, Padding and Predictions: A Deeper Look at Position Information in CNNs,” en, *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3889–3910, Sep. 2024, ISSN: 1573-1405. DOI: 10.1007/s11263-024-02069-9. Accessed: Jan. 9, 2025. [Online]. Available: <https://doi.org/10.1007/s11263-024-02069-9>.
- [258] B. Nam, J. Y. Kim, I. Y. Kim, and B. H. Cho, “Selective Prediction With Long Short-term Memory Using Unit-Wise Batch Standardization for Time Series Health Data Sets: Algorithm Development and Validation,” *JMIR Medical Informatics*, vol. 10, no. 3, e30587, Mar. 2022, ISSN: 2291-9694. DOI: 10.2196/30587. Accessed: Aug. 15, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8965672/>.
- [259] C. Duan and W. Ke, “Advanced Stock Price Prediction Using LSTM and Informer Models,” en, *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, vol. 5, no. 1, pp. 141–166, Jun. 2024, ISSN: 3006-4023. DOI: 10.60087/jaigs.v5i1.183. Accessed: Sep. 1, 2025. [Online]. Available: <https://newjaigs.org/index.php/JAIGS/article/view/183>.
- [260] H. K. Kim and M. H. Sunwoo, “An Automated Cardiac Arrhythmia Classification Network for 45 Arrhythmia Classes Using 12-Lead Electrocardiogram,” *IEEE Access*, vol. 12, pp. 44 527–44 538, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3380892. Accessed: Sep. 1, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10478009>.
- [261] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. Accessed: Jul. 16, 2025. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [262] G. Feretzakis et al., “Integrating Shapley Values into Machine Learning Techniques for Enhanced Predictions of Hospital Admissions,” en, *Applied Sciences*, vol. 14, no. 13, p. 5925, Jan. 2024, Number: 13 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app14135925. Accessed: Jul. 16, 2025. [Online]. Available: <https://www.mdpi.com/2076-3417/14/13/5925>.
- [263] K. Teo et al., “Current Trends in Readmission Prediction: An Overview of Approaches,” en, *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 11 117–11 134, Aug. 2023, ISSN: 2191-4281. DOI: 10.1007/s13369-021-06040-5. Accessed: Jan. 10, 2025. [Online]. Available: <https://doi.org/10.1007/s13369-021-06040-5>.

-
- [264] N. Niako, J. D. Melgarejo, G. E. Maestre, and K. P. Vatcheva, "Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM," en, *BMC Medical Research Methodology*, vol. 24, no. 1, p. 320, Dec. 2024, ISSN: 1471-2288. DOI: 10.1186/s12874-024-02448-3. Accessed: Sep. 1, 2025. [Online]. Available: <https://doi.org/10.1186/s12874-024-02448-3>.
- [265] L.-m. Wang, J.-x. Chen, Y. Pei, X. Zhao, H.-t. Cui, and H.-z. Cui, "Feature Selection and Prediction of Sub-health State Using SVM-RFE," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 3, Oct. 2010, pp. 199–202. DOI: 10.1109/AICI.2010.280. Accessed: Apr. 11, 2024. [Online]. Available: <https://ieeexplore.ieee.org/document/5655277>.
- [266] F. Cabitza and A. Campagner, "The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies," *International Journal of Medical Informatics*, vol. 153, p. 104510, Sep. 2021, ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2021.104510. Accessed: Sep. 9, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1386505621001362>.
- [267] M. Calvert, M. Brundage, P. B. Jacobsen, H. J. Schünemann, and F. Efficace, "The CONSORT Patient-Reported Outcome (PRO) extension: Implications for clinical trials and practice," *Health and Quality of Life Outcomes*, vol. 11, no. 1, p. 184, Oct. 2013, ISSN: 1477-7525. DOI: 10.1186/1477-7525-11-184. Accessed: Sep. 19, 2025. [Online]. Available: <https://doi.org/10.1186/1477-7525-11-184>.
- [268] O. L. Aiyegbusi et al., "Considerations for patient and public involvement and engagement in health research," eng, *Nature Medicine*, vol. 29, no. 8, pp. 1922–1929, Aug. 2023, ISSN: 1546-170X. DOI: 10.1038/s41591-023-02445-x.
- [269] T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, "Deep learning: Systematic review, models, challenges, and research directions," en, *Neural Computing and Applications*, vol. 35, no. 31, pp. 23103–23124, Nov. 2023, ISSN: 1433-3058. DOI: 10.1007/s00521-023-08957-4. Accessed: Sep. 10, 2025. [Online]. Available: <https://doi.org/10.1007/s00521-023-08957-4>.
- [270] X. Huang and J. Marques-Silva, "On the failings of Shapley values for explainability," *International Journal of Approximate Reasoning*, Synergies between Machine Learning and Reasoning, vol. 171, p. 109112, Aug. 2024, ISSN: 0888-613X. DOI: 10.1016/j.ijar.2023.109112. Accessed: Sep. 22, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0888613X23002438>.
- [271] J. Bøttern, T. B. Stage, and A.-C. D. Dunvald, "Sex, racial, and ethnic diversity in clinical trials," *Clinical and Translational Science*, vol. 16, no. 6, pp. 937–945, Mar. 2023, ISSN: 1752-8054. DOI: 10.1111/cts.13513. Accessed: Nov. 20, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10264921/>.

-
- [272] J. Y. Kim, M. Florez, E. Botto, X. Belgrave, C. Grace, and K. Getz, “The influence of socioeconomic status on individual attitudes and experience with clinical trials,” en, *Communications Medicine*, vol. 4, no. 1, p. 172, Sep. 2024, Publisher: Nature Publishing Group, ISSN: 2730-664X. DOI: 10.1038/s43856-024-00586-9. Accessed: Nov. 21, 2025. [Online]. Available: <https://www.nature.com/articles/s43856-024-00586-9>.
- [273] R. Gouveia, V. T. Cruz, and L. Almeida, “Sociodemographic and psychological characteristics influencing patients’ willingness to participate in clinical trials,” *BMJ Open Quality*, vol. 11, no. 4, e002044, Oct. 2022, ISSN: 2399-6641. DOI: 10.1136/bmj-oq-2022-002044. Accessed: Nov. 20, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9628668/>.
- [274] J. E. Alderman et al., “Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations,” English, *The Lancet Digital Health*, vol. 7, no. 1, e64–e88, Jan. 2025, Publisher: Elsevier, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(24)00224-3. Accessed: Dec. 16, 2025. [Online]. Available: [https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(24\)00224-3/fulltext?tpcc=NL_Marketing](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00224-3/fulltext?tpcc=NL_Marketing).
- [275] P. Kokol, M. Kokol, and S. Zagoranski, “Machine learning on small size samples: A synthetic knowledge synthesis,” *Science Progress*, vol. 105, no. 1, p. 00368504211029777, Feb. 2022, ISSN: 0036-8504. DOI: 10.1177/00368504211029777. Accessed: Sep. 11, 2025. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10358596/>.
- [276] A. S. George, A. S. H. George, and A. S. G. Martin, “The Environmental Impact of AI: A Case Study of Water Consumption by Chat GPT,” en, *Partners Universal International Innovation Journal*, vol. 1, no. 2, pp. 97–104, Apr. 2023, ISSN: 2583-9675. DOI: 10.5281/zenodo.7855594. Accessed: Sep. 11, 2025. [Online]. Available: <http://www.puiij.com/index.php/research/article/view/39>.
- [277] D. of Health and Social Care, *10 Year Health Plan for England: Fit for the future*, en, Jul. 2025. Accessed: Dec. 16, 2025. [Online]. Available: <https://www.gov.uk/government/publications/10-year-health-plan-for-england-fit-for-the-future>.
- [278] W. Xuyi, H. Seow, and R. Sutradhar, “Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer,” en, *Cancer Medicine*, vol. 10, no. 3, pp. 989–998, 2021, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cam4.3685>, ISSN: 2045-7634. DOI: 10.1002/cam4.3685. Accessed: Apr. 3, 2024. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cam4.3685>.
- [279] A. Pfob, B. J. Mehrara, J. A. Nelson, E. G. Wilkins, A. L. Pusic, and C. Sidey-Gibbons, “Towards Patient-centered Decision-making in Breast Cancer Surgery: Machine Learning to Predict Individual Patient-reported Outcomes at 1-year Follow-up,” en-US, *Annals of Surgery*, vol. 277, no. 1, e144, Jan. 2023, ISSN: 0003-4932. DOI: 10.1097/SLA.0000000000004862. Accessed: Apr. 11, 2024. [Online]. Available: <https://journals.lww.com/>

annalsofsurgery/fulltext/2023/01000/towards_patient_centered_decision_making_in_breast.47.aspx.

- [280] D. Verma, K. Bach, and P. J. Mork, “External validation of prediction models for patient-reported outcome measurements collected using the selfBACK mobile app,” eng, *International Journal of Medical Informatics*, vol. 170, p. 104936, Feb. 2023, ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2022.104936.

Appendix

8.1 Paper 1

REVIEW

Open Access



Using artificial intelligence to predict patient outcomes from patient-reported outcome measures: a scoping review

Zuzanna Wójcik^{1*} , Vania Dimitrova² , Lorraine Warrington³ , Galina Velikova^{3,4}  and Kate Absolom^{3,5} 

Abstract

Purpose This scoping review aims to identify and summarise artificial intelligence (AI) methods applied to patient-reported outcome measures (PROMs) for prediction of patient outcomes, such as survival, quality of life, or treatment decisions.

Introduction AI models have been successfully applied to predict outcomes for patients using mainly clinically focused data. However, systematic guidance for utilising AI and PROMs for patient outcome predictions is lacking. This leads to inconsistency of model development and evaluation, limited practical implications, and poor translation to clinical practice.

Materials and methods This review was conducted across Web of Science, IEEE Xplore, ACM, Digital Library, Cochrane Central Register of Controlled Trials, Medline and Embase databases. Adapted search terms identified published research using AI models with patient-reported data for outcome predictions. Papers using PROMs data as input variables in AI models for prediction of patient outcomes were included.

Results Three thousand and seventy-seven records were screened, 94 of which were included in the analysis. AI models applied to PROMs data for outcome predictions are most commonly used in orthopaedics and oncology. Poor reporting of model hyperparameters and inconsistent techniques of handling class imbalance and missingness in data were found. The absence of external model validation, participants' ethnicity information and stakeholders involvement was common.

Conclusion The results highlight inconsistencies in conducting and reporting of AI research involving PROMs in patients' outcomes predictions, which reduces the reproducibility of the studies. Recommendations for external validation and stakeholders' involvement are given to increase the opportunities for applying AI models in clinical practice.

*Correspondence:

Zuzanna Wójcik
sczw@leeds.ac.uk

¹ UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care, University of Leeds, Leeds, UK

² School of Computer Science, University of Leeds, Leeds, UK

³ Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK

⁴ Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, Leeds, UK

⁵ Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

Introduction

Artificial Intelligence (AI) is a field of computer science and engineering which uses computer systems able to mimic intelligent behaviour [1]. AI is known to have potential to improve the effectiveness, accessibility and accuracy of screening, diagnosis and treatment in many areas of health [2, 3]. AI models predicting patient outcomes can achieve high performance, and as a result aid clinical decisions and improve quality of healthcare [3]. AI has been applied to various data types in medicine,



using mainly clinical data, such as diagnostic images, genetic data, or brain activity data [4].

While there is a growing attention at patient-reported data in clinical practice and some attempts to use AI models on such data exist [5], systematic guidance on how to apply AI on patient-reported data for outcome predictions is lacking. Patient-reported data can be collected using patient-reported outcome measures (PROMs). These are questionnaires which measure patients' perception on their health status, without being influenced by clinical opinion [6]. PROMs data can be either standardised and validated tools designed to capture patients' reports, or any other forms of symptom and quality of life measures [7]. For instance, mobile applications for PROMs collection, have been widely used in healthcare and have potential to improve the quality and personalisation of patient care [8]. The recent systematic evaluation of PROMs in clinical trials of AI health technologies has shown that patients' perspective is central even in novel technological advancements [9].

Unfortunately, the complexity of PROMs data and limited universal guidelines for AI use in healthcare research [10], can lead to inconsistent reporting of study design and evaluation [11]. Furthermore, studies often lack reproducibility, external validity [12], and generalisability of the results to the clinical context [10]. Inadequate and inconsistent selection of patient-reported input data also introduces a challenge to useful application of patient-centred AI models in healthcare [13]. Additionally, there is a lack of patient and clinician involvement in the process of study design, which plays an important role in addressing bias in AI research for healthcare [14].

There are existing literature reviews exploring AI models applied on PROMs data. For example, a scoping review from 2021 investigated PROMs as standalone input variables in models, however, they did not explore reproducibility and clinical adoption of studies. Moreover, only 2 medically oriented literature databases were searched, while databases from engineering and computer science backgrounds were not considered [5]. Other existing reviews focused on specific healthcare domains (e.g. oncology) and did not investigate all potential applications of using AI and PROMs data. [15–17].

This review aims to address the gap in the literature by investigating AI models used in primary studies for predicting patient outcomes using PROMs. It focuses on methodological rigour of conducting, evaluating and reporting AI research including PROMs as input data. It highlights the importance of ensuring standardised dataset description and justification for chosen methods of model development and evaluation, focusing on clinical relevance. Recommendations for engaging stakeholders, including patients, are suggested.

Materials and methods

The methodology of this scoping review was based on the Joanna Briggs Institute (JBI) guidance [18]. The review protocol is available on Open Science Framework [19]. The completed PRISMA checklist for scoping reviews [20] is added in the Supplementary Materials Fig. 1 and 2.

Search strategy

The databases used to search for relevant papers were: Web of Science, IEEE Xplore, ACM Digital Library, Cochrane Central Register of Controlled Trials, Medline and Embase. These databases were selected to include the variety of fields publishing papers on AI in medicine, covering both medical and engineering aspects. The keywords adapted to each database are listed in Supplementary Materials Table 1. Initially, the limited search of Web of Science and Medline was conducted to analyse and approve the keywords. The finalised search of all the records across all the databases was completed on the 7th November 2023. The reference lists of all relevant reports were also screened. All studies identified through the search strategy were exported to Endnote citation management system.

Inclusion and exclusion criteria

The inclusion and exclusion criteria followed the Population/Concept/Context (PCC) framework [18] and are described in Table 1. The participants in the papers included in the review were patients, whose symptoms and quality of life data were recorded using various PROMs. These can include mobile applications, or surveys completed either online or in a clinic. The type of data can be collected through either validated and widely used PROMs, or any other patient self-reports. Papers reporting the use of PROMs data as both a predictive and predicted variable were included in the study. If PROMs data were only used as a predicted variable, and not included as inputs, the reference was excluded. The concept was the methods of AI used for predictions of the patients' outcomes. Papers that explicitly mentioned use of AI or Machine Learning methods were included. Any papers using AI models for purposes other than prediction were excluded. Papers that reported prediction of patient outcomes in the healthcare context were included in the analysis. These outcomes should belong to the categories of patient-related outcomes identified by Kersting et al. (2020) [21], presented in Table 1. The broad understanding of healthcare context allowed focusing on the AI used for various medical reasons.

Review process

All duplicates found in the databases were removed automatically in Endnote. Titles and abstracts of the

Table 1 Inclusion and Exclusion criteria for the study selection

Category	Inclusion criteria	Exclusion criteria
Input variables	Data reported by patients using standardised PROMs; electronic data collection designed for self-reporting of symptoms or clinical outcomes. These can be used with combination of different non-patient reported data	Only not patient-reported data, e.g., data reported through a clinician, recorded, written down during an appointment, data reported on online forums/social media, or data from physiological measurements
Output variables	Any data describing patient outcomes: sleep behaviour, coping and self-efficacy, healthcare utilisation, body image perception, function, communication skills, reliability of diagnosis and therapy, optimal support, confidence in therapy, satisfaction, cognitive performance, treatment decision, disease control, daily activities, reoperation, mental health, quality of life, mobility, co-morbidities, pain, survival, adverse events, and symptoms	Output that does not relate to health or healthcare, or papers which did not aim to predict outcomes
Models	Machine learning or deep learning predictive models	Statistical models (e.g., only regression analysis)
Paper type	Primary research reported in English language	Abstracts only, theses, dissertations, letters to editors, guidelines, commentaries, introductions, papers published not in English, or review papers

papers were screened by a researcher and re-selected based on inclusion and exclusion criteria, presented in Table 1. Full texts of articles admitted to the study were assessed against the inclusion and exclusion criteria again. The researcher’s approach was validated through a second reviewer, who repeated scanning through 10% of abstracts, selected full texts and compared their results with the first reviewer. The validation showed high consistency between the reviewers’ decisions, as out of 218 validated papers, 184 (84.4%) were consistently selected or rejected. Therefore, no further validation was performed.

Data extraction and analysis

The data was extracted from all papers selected for this review. Extracted and summarised information for each included paper is presented in Supplementary Materials Tables 1 and 2. A second reviewer extracted data from 10% of admitted papers for the purpose of validation, and the extracted information was compared and agreed between the reviewers. The summary of data was reported in tabular form in Excel spreadsheet and presented in a narrative form in this review.

The extracted and analysed data included:

- Study characteristics (country and year of publication, healthcare domain, input PROMs variables used, types of PROMs, output variable types, and sample sizes)
- Data pre-processing (missingness in the datasets, missing data imputation techniques, class distribution, techniques for handling class imbalance)

- Model development (types of AI models used, frequency of AI models used, AI techniques for addressing temporality in data, hyperparameter tuning)
- Model evaluation (performance metrics used, variable importance, best-performing AI models)
- Clinical relevance and adoption (patients and clinicians involvement in the study design, validation and deployment stage of research, reporting of sociodemographic information)

Results

Out of 3077 records screened, 94 were selected for analysis in this review. PRISMA diagram [20] (Fig. 1) illustrates the process of paper selection. The reasons for paper exclusions were: no full-text available (38.1%), no PROMs used as input variables (35%), or no AI models used (12%), or methods did not aim to predict patient outcomes (8.75%).

Study characteristics

Among the identified studies, 33 (35%) were conducted in USA, 31 (33%) in Europe, 6 (6%) in Canada, 15 (16%) in Asia, 1 (1%) in South America, 1 (1%) in New Zealand, and 1 (1%) in Turkey. Six (6%) studies were conducted internationally (USA and Canada ($n = 2$, 2%); UK and USA ($n = 2$, 2%); Canada and Sweden ($n = 1$, 1%); Europe, US, Australia and Israel ($n = 1$, 1%)). Identified papers focused on orthopedics ($n = 38$, 40%), oncology ($n = 22$, 23%), mental health ($n = 17$, 18%), respiratory ($n = 8$, 9%), neurology ($n = 4$, 4%) and other domains ($n = 5$, 5%), which appeared only once: hearing, endometriosis, palliative care, sub-health state, and cardiovascular. The studies were published between 2010 and 2023 (Fig. 2). The

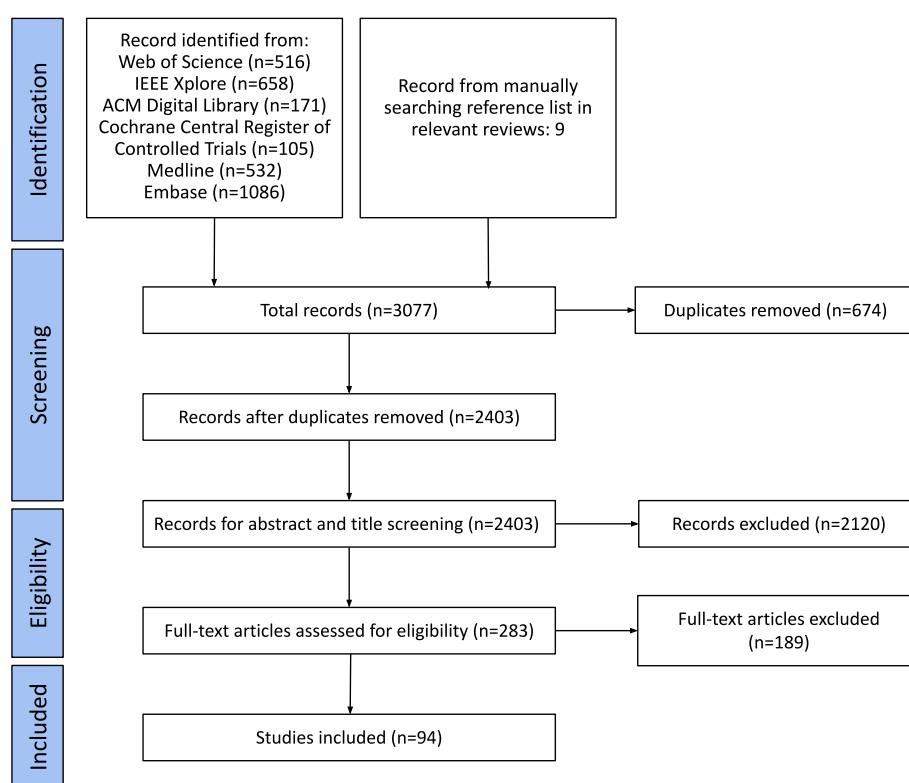


Fig. 1 PRISMA flow diagram

data were obtained either from existing registry/database ($n = 44$, 47%), or pre-existing or current research studies ($n = 47$, 50%, not reported: $n = 3$, 3%). The self-reported input variables were combined with clinical and demographic data ($n = 63$, 67%), only demographic data ($n = 14$, 15%), only clinical data ($n = 3$, 3%), or other types of data ($n = 4$, 4%), such as wearable, electroencephalographic, bio-mechanical, or family data. Ten studies used only self-reported data for predictions. Most papers ($n = 63$, 67%) were predicting self-reported outcomes, 14 (15%) of which were Minimally Clinically Important Differences (MCID) between pre- and post-clinical event data collection. Other papers used either only objectively measured outcome ($n = 28$, 30%), or a combination of self-reported and objective outcomes ($n = 3$, 3%). Sample sizes of the papers varied from 20 to 1,434,868 (mean = 25,888, median = 1022, 1st quartile = 429.75, 3rd quartile = 2879.75). The quartiles do not indicate clear boundaries between the data, as there are small differences in the sample sizes. Hence, a boundary-based approach was followed instead of quartile-based: very small (< 300), small (300–700), medium (701–2000), large (2001–20000), and very large ($> 20,000$), as presented in Fig. 3. The vast majority of studies ($n = 69$, 73%) used condition-specific PROMs, such as orthopedic-specific Knee Injury

and Osteoarthritis Outcome Score (KOOS) [22] or cancer-specific EORTC Core Quality of Life Questionnaire (QLQ-C30) [23]. In 31 papers (33%) condition-specific measures with generic questionnaires, for example Euro-Qol- 5D (EQ- 5D) [24]. Twelve papers (13%) used generic measures only. Out of 81 (86%) papers that reported the types of questionnaires used, 18 (22%) focused on physical health, 11 (14%) on mental health, and 52 (64%) on both.

Data pre-processing

Thirty papers did not report the missingness in the dataset. Therefore, it is uncertain if the datasets in these studies did not have any missing data, or if missing data were not disclosed. Out of 64 papers (68%) which reported missing data, only 1 (2%) stated that there was no missingness in the dataset. Ten papers (16%) which reported having missing data did not report how the missingness was handled or addressed. Out of all papers, only 53 (56%) reported the technique for data imputation (Fig. 4c). The 2 most common techniques were complete case analysis ($n = 16$, 30%), and mean/median/mode imputation ($n = 15$, 28%). Most papers ($N = 89$, 95%) used classification as a prediction method. Fourteen (16%) of these did not provide any information about the class

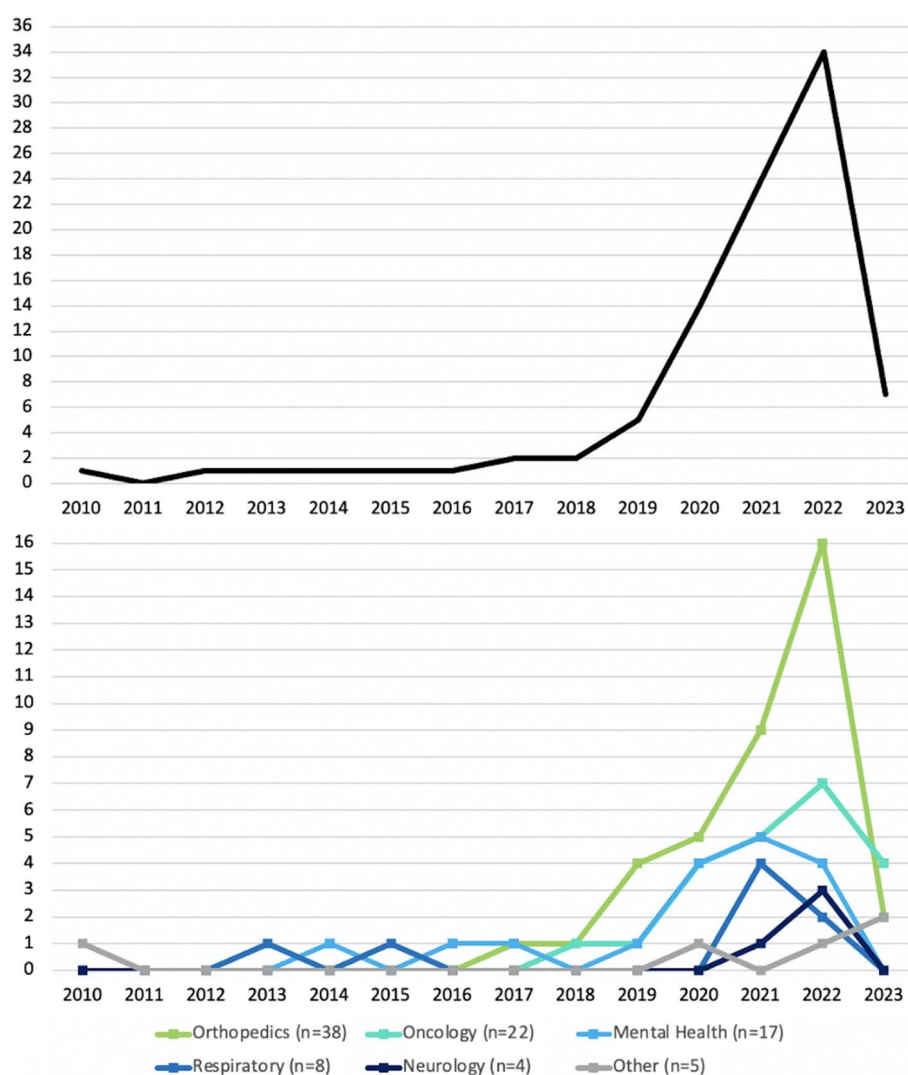


Fig. 2 Year of publication of all 94 studies (top figure) and studies based on health domain (bottom figure)

distribution. All papers which reported class distribution ($n = 75$, 80%) performed binary classification. Out of these papers, only 11 (15%) had balanced classes (maximum imbalance ratio of 60:40 between the minority and majority class [25]). Sixty-four papers (68%) used dataset with imbalanced classes, 29 (45%) of which did not mention the class imbalance problem. Thirty-five papers (55%) acknowledged the issue but 13 (37%) of them left the data imbalanced. In total, 22 papers (23%) reported the need for balancing the classes, but there was inconsistency in the methods across papers (Fig. 4).

Model development

Most papers ($n = 84$, 89%) used multiple AI models for outcomes prediction. Forty papers (43%) used only traditional machine learning (ML) models, 5 (5%) only deep learning (DL) models, and 49 (52%) both ML and

DL. The most frequently used models were regression models ($n = 61$, 65%), including linear, logistic, ridge and LASSO regressions; boosting methods ($n = 53$, 56%), including adaptive boosting, extreme gradient boosting and gradient boosting machine; random forest ($n = 50$, 53%); artificial neural network ($n = 43$, 46%), including single-or multi-layer perceptrons; and support vector machine ($n = 39$, 41%) (Fig. 5). Most studies ($n = 74$, 81%) applied AI models on data recorded in one time-point. The remaining studies trained their models on data collected in multiple time-points (Table 2). Out of these, 3 studies (3%) reported using models that process the temporal dependencies in the data, such as long-short term memory (LSTM) model [26, 27], and recurrent neural network with gated recurrent units (GRU) [28]. Nine studies (12%) considered temporality through coding it in the feature sets, and 5 papers (7%) did not address

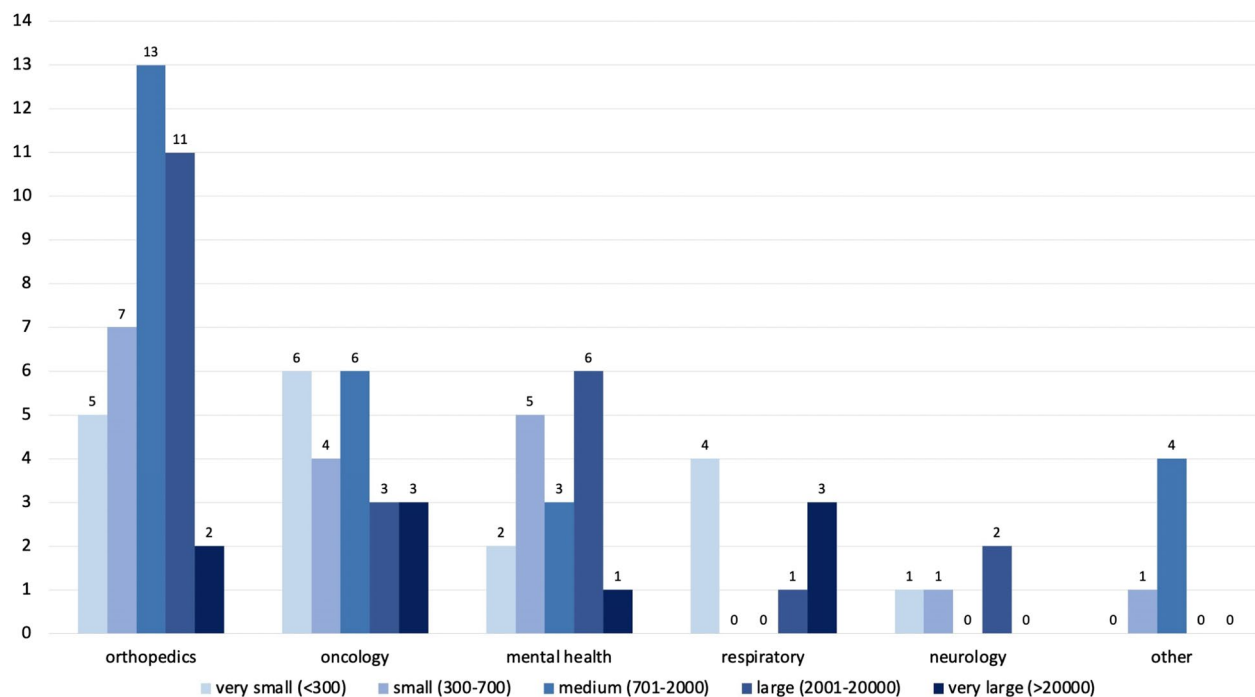


Fig. 3 Number of papers categorised based on sample sizes in each healthcare domain

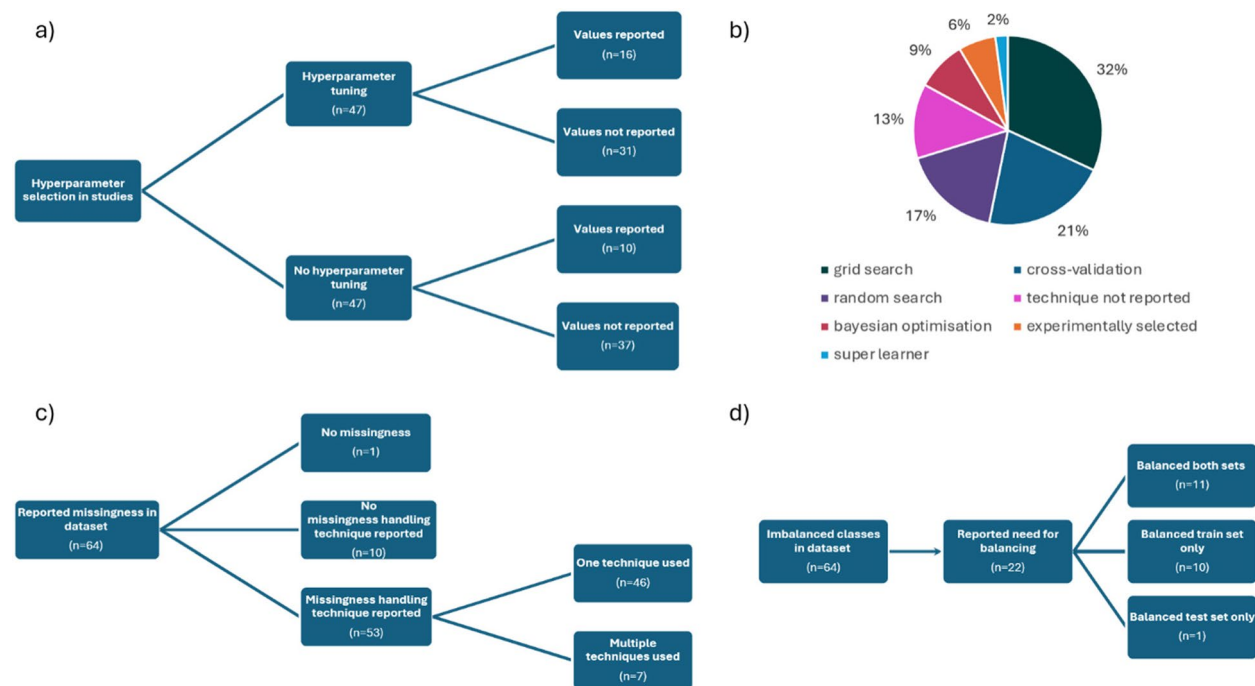


Fig. 4 Reporting of pre-processing and model development methods in the studies. Sub-figure **a**) Frequency of hyperparameter tuning and values reporting. Sub-figure **b**) Proportion of hyperparameter tuning techniques. Sub-figure **c**) Missingness reporting and imputation in papers. Sub-figure **d**) Handling class imbalance in studies

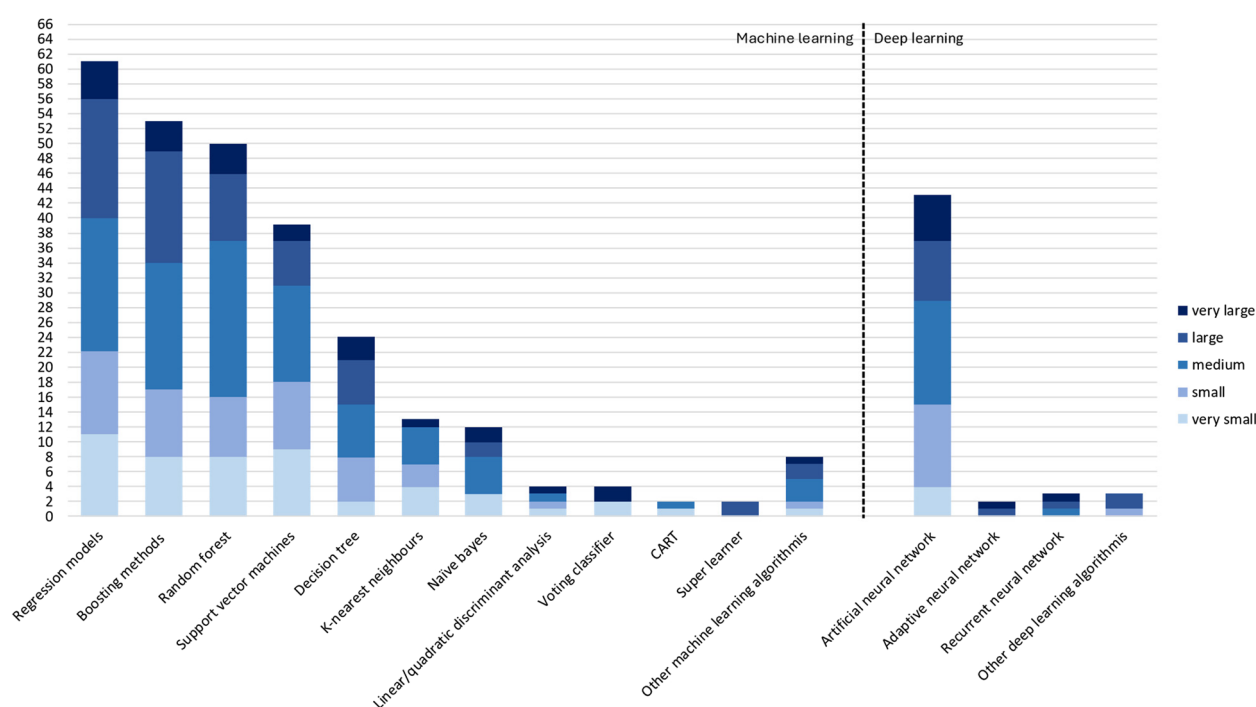


Fig. 5 Frequency of algorithms used on datasets with very small (fewer than 300), small (300–700), medium (701–2000), large (2001–20000) and very large (more than 20,000) sample size

temporality at all (Table 3). Half of the papers in this review ($n = 47$, 50%) reported performing hyperparameter tuning, and out of these only 16 (34%) reported used hyperparameters (Fig. 4a)).

Model evaluation

The evaluation metrics varied across the studies. Area under the curve (AUC) was most commonly used ($n = 60$, 64%), and 32 (53%) of papers used this value to assess model performance with imbalanced classes. Other frequently used performance metrics were recall, also known as sensitivity ($n = 44$, 47%), accuracy ($n = 43$, 46%), and specificity ($n = 34$, 36%). The majority of the studies used multiple performance metrics ($n = 83$, 88%). Variable importance analysis was performed by 64 studies (68%) 61 of which (95%) reported PROMs data being valuable for prediction. Seventy-nine papers (84%) provided information on the best performing model. Regression models were the most frequently selected as best-performing algorithms ($n = 24$, 30%), followed by boosting methods ($n = 20$, 25%), random forest ($n = 10$, 13%) and neural network ($n = 10$, 13%).

Clinical relevance and adoption

No studies reported that the developed methods had been applied in the clinical practice. Although we acknowledge, that in such multidisciplinary research

clinicians are generally involved in the study design, only 3 papers (3%) explicitly mentioned how clinicians contributed to the model development. They helped selecting input variables [30, 111], or creating a testing set [100]. No papers mentioned patients involvement in the model development or any part of the study design. The majority of papers reported age ($n = 63$, 67%) and gender ($n = 60$, 64%) of study participants and only 24 (26%) studies reported ethnicity. Papers were classified into 3 different categories, inspired by a previously conducted scoping review [118]: internal validation (one source of data used for training and validation, including cross-validation or holdout sample for validation on unseen data from the same dataset), external validation (the model developed on one dataset and then tested/validated on a completely new (i.e. external) dataset) or deployment ("integrated into a prototype application, and evaluated for its feasibility in clinical workflows"[118]). Based on these definitions, 81 papers (86%) were in the internal validation stage, 10 (11%) completed external validation, and 3 papers (3%) were in the deployment stage.

Discussion

This scoping review aimed to identify AI methods used on PROMs data to predict patient outcomes. The analysis of 94 papers allowed the exploration of algorithms applied on complex patient-reported data and revealed

Table 2 Machine learning and deep learning models used on data collected in one and multiple timepoints, ordered by number of publications

Machine Learning (ML)	One Timepoint	Multiple Timepoints
Regression ($n = 61$)	[29–39] [40–50] [40, 51–60] [61–71] [72–75] [76]	[77–86] [28, 87, 88]
Boosting ($n = 53$)	[31, 32, 36–40, 42, 44, 89, 90] [45, 48, 49, 51–53, 91–94] [40, 56, 58, 60, 62–64, 95–97] [2, 33, 66, 69, 72, 73, 75, 98–100] [70, 74, 101] [76] [102]	[28, 77–79, 81–84, 103, 104]
Random Forest ($n = 50$)	[29–33, 36, 41, 42, 44, 89, 105] [45–48, 50–53, 56, 92, 106] [58, 60, 61, 64, 65, 69, 75, 95, 100, 107, 108] [62, 63, 71, 73, 74, 97, 101]	[26, 28, 77, 78, 80, 87, 88, 104, 109, 110]
Support Vector Machine ($n = 39$)	[29, 31, 32, 34, 37, 41–43, 45, 46, 48] [26, 50–52, 56, 57, 92, 96, 111, 112] [61, 62, 64, 66, 71, 74, 89, 107, 108] [102]	[78, 81, 83, 84, 87, 88, 103, 104]
Decision Tree ($n = 24$)	[37, 39, 41, 57, 92–95, 111, 112] [2, 58, 63, 74, 96, 98, 100, 101, 108]	[77, 79, 84, 103, 109]
K-Nearest-Neighbours ($n = 13$)	[34, 36, 37, 39, 48, 56, 61, 74, 113]	[26, 83, 87, 104]
Naïve Bayes ($n = 12$)	[36, 44, 50, 57, 70, 73, 74, 94, 108]	[26, 104, 109]
Voting Classifier ($n = 4$)	[37, 44]	[28, 82]
Discriminant Analysis ($n = 4$)	[34, 38, 94, 108]	None reported
Classification and Regression Tree ($n = 2$)	[34, 46]	None reported
Super Learner ($n = 2$)	[62, 69]	None reported
Other ML Methods ($n = 8$)	Wide and Deep [49] Stochastic Gradient Descent [61] Bagging [63] Bayesian Updating Algorithm [66] Graphical Gaussian Model [67] Multivariate Adaptive Regression Spline [41]	Hierarchical Gaussian Process [85] Autoregressive Integrated Moving Average [26]
Deep Learning (DL)	One Timepoint	Multiple Timepoints
Multilayer Perceptron ($n = 43$)	[29, 30, 35–37, 39–42, 45, 90] [48, 52–54, 56, 92, 94, 111, 112, 114] [40, 57, 59, 61, 64, 66, 68, 96, 107, 115] [72–75] [76]	[26, 28, 79, 83, 84, 86, 104, 109, 116]
Recurrent Neural Network (RNN) ($n = 3$)	None reported	Long-Short Term Memory [26, 27] RNN with Gated Recurrent Units [28]
Other DL Methods ($n = 4$)	Adaptive Neural Network [36] Stacking Algorithm [57] Bayesian Network Model [117]	Adaptive Neural Network [27]

In the square brackets we list the number of the cited paper, according to the reference list

the opportunities, challenges and best practice recommendations for AI medical research involving PROMs. The main findings suggest the variety of data types and evaluation metrics used, as well as inconsistencies in data pre-processing and model development design and reporting.

Study characteristics

Due to fragmented data collection of PROMs, incorporating them into AI systems is very challenging [119]. Therefore, the majority of papers in this study have small sample size. In orthopedics settings PROMs have been increasingly collected as a part of routine care [120], which explains the large proportion of orthopedics papers with medium-to-large datasets. The large sample size was common in mental health papers, as

Table 3 Methods of addressing temporality in time-series data

Level of addressing temporality	Method of addressing temporality	Description of the method	Sample size	Frequency
Addressed by the model	Recurrent Neural Network (RNN)	Data transformed to 3D array and fed in the LSTM model [26] Events encoded by Adaptive Net, pooled by LSTM model [27]	823 9,500	Weekly Irregular ^a
		RNN with GRU, considering each treatment as timestep [28]	105,129	Irregular ^a
Addressed in the features	Measured change	Change of measurement from baseline [84]	245	Every 90 days
		Change in mean measurements from baseline [88]	31,700	Daily
		Mean daily change from the 24-h baseline period [87]	116	Twice a day
		Change in symptom severity from previous report [78]	34	Weekly
		Variable indicated if a report is followed by exacerbation event [80] Occurrence of symptom in any day of a time window [85]	2,374 182,991	Daily Daily (3 days)
	Binary outcome	Dichotomised score one week following the prediction date [104]	210	Weekly
Not considered	Feature for each timeline	Score added as an input feature at every measurement [82] Created a timeline of best overall responses (BORs) [103]	83 31	Weekly Weekly
		Treated the 2- and 8-week measures as if assessed at baseline [81] Three models that used 7, 14, and 21 days as inputs [116]	1,003 20	3 time-points 3 time-points
	Selected 1 value for analysis	If patient had multiple follow-up events, the first was chosen [77] The assessment with the highest overall score was used [79]	494 11,761	Irregular ^a Bi-weekly
		Score was updated at each assessment [86]	212,615	Irregular ^a

^a Irregular measurements indicate that the reports were completed at any clinical event that occurred

Three papers which used time-series data did not report how the temporality was addressed [83, 109, 110], and are not included in this table

mental health screening and diagnostic tools are usually based on PROMs, and there is a long-standing history of using such tools [121]. Contrary to orthopedics and mental health settings, PROMs collection in other healthcare domains is very limited. The respiratory datasets were mainly very small or very large. The papers with very large sample size were predicting outcomes related to COVID-19 pandemic, where mobile applications collecting PROMs became more common [122]. Most of the studies analysed data collected specifically for research studies, rather than in clinical practice, which might introduce biases related to inclusion and exclusion criteria. Due to inconsistent PROMs questionnaires used across different studies, the comparison of results is limited. Therefore, using standardised and validated measures can help explore the overall predictive value of PROMs. The peak in the use of AI methods for all domains was between 2021 and 2022. This recent increase is compatible with a scoping review on AI in

healthcare, where 71% of studies were published between 2020 and 2022 [123].

Data pre-processing

Missing data in AI research is an important aspect to investigate, as it can lead to various biases [124]. Therefore, the inconsistencies in reporting data quality in the analysed studies are concerning. The justification of using data imputation techniques was also poor, whilst most commonly used techniques (complete case analysis and mean/median/mode imputations) can frequently cause bias [124]. Only a small number of papers used KNN-based imputation, which can reach the accuracy of complete data with a low performance difference [125]. Studies applying AI methods on PROMs should ensure that missing data are reported and any imputation methods are justified [124]. Another inconsistency between the papers was caused by various methods for handling class imbalance in classification tasks. Papers

which reported class imbalance often did not attempt to balance the data, which prevents models from appropriate learning from the training set. Furthermore, most of these papers used AUC as a performance metric, which require the balanced setting to avoid bias [126]. The papers which reported balancing data have also done it inconsistently and without justification. Balancing data prior to train and test split can cause issues in model validation, as real data is never perfectly balanced. Therefore, it is important to evaluate model performance on test set unaffected by sampling methods [127]. The choice of performance metrics should also be justified and able to uncover potential bias caused by class imbalance (for example balanced accuracy and F1 score, instead of accuracy and AUC).

Model development

The studies in this review reported model development process inconsistently, with majority of studies missing model hyperparameters reporting. According to Jha et al. (2023)[128], it should be “the ethical obligation” to document all stages of model development that are essential for the reproducibility of results. Therefore, model hyperparameters and their optimisation technique should always be reported and justified. The missingness in data was also handled and reported inconsistently, which is an important step for reproducibility as well. The lack of large PROMs datasets prevents applying deep learning methods, which can be extremely useful in capturing patterns in high-dimensional data or dependencies that other algorithms can’t capture [129]. Only the simple “vanilla” neural network was applied more often than some of the basic ML models. Studies which collected data in multiple timepoints often did not address the temporality at all or analysed time-series data through data pre-processing strategies and conventional ML models. Only 3 papers used DL models that are appropriate for temporal processing. These are for example LSTM or GRU methods. Most papers in this review chose various ways to include temporal information through feature engineering, as described in Table 3. Nevertheless, DL models have been more successful in accurate predictions of patients’ outcomes when applied to time-series data, as they are able to process more complex dependencies in high dimensionality and temporality of medical data [130].

Model evaluation

Most studies used multiple evaluation metrics, which allow between-studies comparisons and in-depth analysis of model performance. Variable importance analysis was also commonly conducted, which supports the explainability of AI models [131]. Furthermore, the studies used

multiple models, which allowed them to select the best-performing one. The analysis of these showed that most common models were rarely selected as best models (e.g., random forest was selected as the best model only 20% of the times). However, voting classifier was used only 4 times, but selected as the best model 3 times. This suggest that further studies should perhaps pay more attention to models that are used less-frequently, which have a potential to perform better.

Clinical relevance and adoption

Limited reporting of clinician and patient engagement in the study development is of concern. This process is known as crucial for ensuring a patient-centred research and feasibility of the studies [132]. Involving stakeholders also can help building trust of the public to AI researchers, and as a result support the implementation of the studied tools in clinical practice [133]. Another issue arising from this review is the lack of external validation of the model performance in the studies, which is an essential step to potential clinical adoption. Assessing the model performance in a different setting may show different model performance which might suggest bias in the original study [118]. This review shows that new models keep being developed to address original problems, without taking the studies further and exploring their potential in the real-world settings. Therefore, validating existing models on external datasets and communicating the design and results with stakeholders should be the next step to support the adoption of AI methods in clinical practice. Furthermore, the majority of papers did not provide any information on the ethnicity of the study participants. Ensuring diverse study population is an essential ethical consideration and lack of ethnic information can further contribute to deepening healthcare inequalities [128].

Strengths and limitations

The main strength of this review is that it identifies AI models applied on self-reported data for predicting patient outcomes in all healthcare domains. The use of 6 different databases from both health and computer science field helped reaching many relevant papers, which might have been omitted by reviews using limited number of databases. This paper also analyses the rigour of model development and evaluation reporting. It focuses on clinical adoption potential, from the perspective of patients and clinicians involvement and ethical consideration of participants’ diversity. The limitations of this study include the possibility of omitting the studies published in language other than English. Only published studies were considered, which might affect the conclusions and further deepen the publication bias. Since the

study focuses on rigour in model development, evaluation, and wider stakeholder engagement, it is important to note, that the results are based on what was reported, and not what was done in the included papers.

Conclusions

The analysis of 94 papers in this scoping review revealed the potential of using PROMs data in AI healthcare research, and inconsistencies in conducting and reporting these studies. It showed the importance of justification of chosen data pre-processing and model development methods, and the involvement of all stakeholders during the study. Our future work will involve applying AI on PROMs data and further explore the potential of time-series patient-reported data for healthcare outcomes predictions. We believe that insights from this paper can inform the rigorous implementation of AI models in clinical practice.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12955-025-02365-z>.

Supplementary Material 1. Figure 1: PRISMA checklist for scoping reviews part 1. Figure 2: PRISMA checklist for scoping reviews part 2. Table 1: Search strategy. Table 2: Study characteristics and pre-processing methods used by studies included in the review. Table 3: Model development and evaluation, including study characteristics of papers included in the review.

Acknowledgements

This work was supported in part by UK Research and Innovation (UKRI) [CDT grant number EP/S024336/1]. The authors would like to thank Shazeeha Masud for the help with data extraction validation, and Professor David Hogg, Dr Amy Downing, and Allan Pang for providing feedback on the initial draft of this paper.

Authors' contributions

Zuzanna Wójcik: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review & Editing, Visualization, Project administration Vania Dimitrova: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Visualization, Supervision, Funding Acquisition Lorraine Warrington: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing—Original Draft, Writing—Review & Editing, Supervision Galina Velikova: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Supervision, Funding Acquisition Kate Absolom: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing—Original Draft, Writing—Review & Editing, Supervision.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

Galina Velikova: Honoraria: Pfizer, Novartis, Eisai, Lilly Advisory boards: Consultancy fees from AstraZeneca, Roche, Novartis, Pfizer, Seagen, Eisai, Sanofi April 2024 = AZ working group (unpaid) Institutional grant: Pfizer.

Received: 30 August 2024 Accepted: 25 March 2025

Published online: 11 April 2025

References

- Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2019;9(4): e1312.
- P. Rajpurkar, J. Yang, N. Dass, V. Vale, A.S. Keller, J. Irvin, and L.M. Williams Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. *JAMA network open*, 3(6), e206653–e206653. 2020
- Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*. 2022;6:228–39.
- Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*. 2017;2:230–43.
- D. Verma, K. Bach, and P. J. Mork, "Application of Machine Learning Methods on Patient Reported Outcome Measurements for Predicting Outcomes: A Literature Review," *Informatics*, vol. 8, p. 56, 2021. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- Kingsley C, Patel S. Patient-reported outcome measures and patient-reported experience measures. *BJA Education*. 2017;17:137–44.
- D. Denny, S. B. Nutakki, K. Alston, and M. Markman, "A review of patient self-reported symptom data with a focus on pain," *Journal of Clinical Oncology*, vol. 32, pp. 79–79, 2014. Publisher: Wolters Kluwer.
- Awad A, Trenfield SJ, Pollard TD, Ong JJ, Elbadawi M, McCoubrey LE, Goyanes A, Gaisford S, Basit AW. Connected healthcare: Improving patient care using digital health technologies. *Adv Drug Deliv Rev*. 2021;178: 113958.
- F. J. Pearce, S. C. Rivera, X. Liu, E. Manna, A. K. Denniston, and M. J. Calvert, "The role of patient-reported outcome measures in trials of artificial intelligence health technologies: a systematic evaluation of ClinicalTrials.gov records (1997–2022)," *The Lancet Digital Health*, vol. 5, pp. e160–e167, 2023. Publisher: Elsevier.
- B. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, and S. Abdullah, "Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector," *Biomedical Materials & Devices* (New York, N.y.), pp. 1–8, 2023.
- S. Jayakumar, V. Sounderajah, P. Normahani, L. Harling, S. R. Markar, H. Ashrafi, and A. Darzi, "Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study," *npj Digital Medicine*, vol. 5, pp. 1–13, 2022. Number: 1 Publisher: Nature Publishing Group.
- Trocin C, Mikalef P, Papamitsiou Z, Conboy K. Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Inf Syst Front*. 2023;25:2139–57.
- S. C. Rivera, X. Liu, S. E. Hughes, H. Dunster, E. Manna, A. K. Denniston, and M. J. Calvert, "Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies," *The Lancet Digital Health*, vol. 5, pp. e168–e173, 2023. Publisher: Elsevier.
- Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns*. 2021;2: 100347.
- Engstrom T, Tanner S, Lee WR, Forbes C, Walker R, Bradford N, Pole JD. Patient reported outcome measure domains and tools used among adolescents and young adults with cancer: A scoping review. *Crit Rev Oncol Hematol*. 2023;181: 103867.
- P. Jayakumar, E. Lin, V. Galea, A. J. Mathew, N. Panda, I. Vetter, and A. B. Haynes, "Digital Phenotyping and Patient-Generated Health Data for Outcome Measurement in Surgical Care: A Scoping Review," *Journal*

- of Personalized Medicine, vol. 10, p. 282, 2020. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
17. S. S. Hoque, S. Ahern, H. E. O'Connell, L. Romero, and R. Ruseckaite, "Comparing patient-reported outcome measures for pain in women with pelvic floor disorders: a scoping review," *The Journal of Pain*, 2023.
18. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*. 2015;13:141–6.
19. Wójcik Z, Dimitrova V, Warrington L, Velikova G, Absolom K. Predicting patient outcomes from self-reported symptom data using Artificial intelligence: a scoping review protocol. OSF: Publisher; 2023.
20. A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, S. Hempel, E. A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M. G. Wilson, C. Garrity, S. Lewin, C. M. Godfrey, M. T. Macdonald, E. V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Tun, calp, and S. E. Straus, "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation," *Annals of Internal Medicine*, vol. 169, pp. 467–473, 2018.
21. Kersting C, Kneer M, Barzel A. Patient-relevant outcomes: what are we talking about? A scoping review to improve conceptual clarity. *BMC Health Serv Res*. 2020;20:596.
22. E. M. Roos, H. P. Roos, L. S. Lohmander, C. Ekdahl, and B. D. Beynon, "Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a Self-Administered Outcome Measure," *Journal of Orthopaedic & Sports Physical Therapy*, vol. 28, pp. 88–96, 1998. Publisher: Journal of Orthopaedic & Sports Physical Therapy.
23. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JC. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85:365–76.
24. Rabin R, de Charro F. EQ-5D: a measure of health status from the Euro-Qol Group. *Ann Med*. 2001;33:337–43.
25. Koyyada SP, Singh TP. A multi stage approach to handle class imbalance: An ensemble method. *Procedia Computer Science*. 2023;218:2666–74.
26. Wang Y, Van Dijk L, Mohamed ASR, Fuller CD, Zhang X, Marai GE, Canahuate G. "Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes," *Proceedings. International Database Engineering and Applications Symposium*. 2021;2021:273–9.
27. M. Kalweit, U. A. Walker, A. Finckh, R. Mu. Iler, G. Kalweit, A. Scherer, J. Boedecker, and T. Hu. gle, "Personalized prediction of disease activity in patients with rheumatoid arthritis using an adaptive deep neural network," *PLoS One*, vol. 16, no. 6, p. e0252289, 2021.
28. R. C. Grant, J. C. He, F. Khan, N. Liu, S. Podolsky, Y. Kaliwal, M. Powis, F. Notta, K. K. W. Chan, M. Ghassemi, S. Gallinger, and M. K. Krzyzanowska, "Machine Learning–Based Early Warning Systems for Acute Care Utilization During Systemic Therapy for Cancer," *Journal of the National Comprehensive Cancer Network*, vol. 21, pp. 1029–1037.e21, 2023. Publisher: National Comprehensive Cancer Network Section: Journal of the National Comprehensive Cancer Network.
29. Klemm C, Uzozike AC, Esposito JG, Harvey MJ, Yeo I, Subih M, Kwon Y-M. The utility of machine learning algorithms for the prediction of patient-reported outcome measures following primary hip and knee total joint arthroplasty. *Arch Orthop Trauma Surg*. 2023;143:2235–45.
30. Tennenhouse LG, Marrie RA, Bernstein CN, Lix LM. Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease. *J Psychosom Res*. 2020;134: 110126.
31. Zhang S, Chen JY, Pang HN, Lo NN, Yeo SJ, Liow MHL. Development and internal validation of machine learning algorithms to predict patient satisfaction after total hip arthroplasty. *Arthroplasty*. 2021;3:33.
32. Zhang S, Lau BPH, Ng YH, Wang X, Chua W. Machine learning algorithms do not outperform preoperative thresholds in predicting clinically meaningful improvements after total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2022;30:2624–30.
33. D. Verma, K. Bach, and P. J. Mork, "Using Automated Feature Selection for Building Case-Based Reasoning Systems: An Example from Patient-Reported Outcome Measurements," in *Artificial Intelligence XXXVIII: 41st SGA International Conference on Artificial Intelligence*, AI 2021, Cambridge, UK, December 14–16, 2021, *Proceedings, (Berlin, Heidelberg)*, pp. 282–295, Springer-Verlag, 2021.
34. Buus AA, Udsen FW, Laugesen B, El-Galaly A, Laursen M, Hejlesen OK. Patient-Reported Outcomes for Function and Pain in Total Knee Arthroplasty Patients. *Nurs Res*. 2022;71:E39–47.
35. V. E. Staartjes, M. P. de Wispelaere, W. P. Vandertop, and M. L. Schr oder, "Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling," *The Spine Journal: Official Journal of the North American Spine Society*, vol. 19, pp. 853–861, 2019.
36. Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med Inform Decis Mak*. 2019;19:3.
37. C. Xu, I. M. Subbiah, S.-C. Lu, A. Pfof, and C. Sidey-Gibbons, "Machine learning models for 180-day mortality prediction of patients with advanced cancer using patient-reported symptom data," *Quality of Life Research*, 2022.
38. A. H. S. Harris, A. C. Kuo, T. R. Bowe, L. Manfredi, N. F. Lalani, and N. J. Giori, "Can Machine Learning Methods Produce Accurate and Easy-to-Use Preoperative Prediction Models of One-Year Improvements in Pain and Functioning After Knee Arthroplasty?," *The Journal of Arthroplasty*, vol. 36, pp. 112–117.e6, 2021. Publisher: Elsevier.
39. A. Siccoli, M. P. d. Wispelaere, M. L. Schr oder, and V. E. Staartjes, "Machine learning–based preoperative predictive analytics for lumbar spinal stenosis," *Neurosurgical Focus*, vol. 46, p. E5, 2019. Publisher: American Association of Neurological Surgeons Section: Neurosurgical Focus.
40. A. Pfof, B. J. Mehrara, J. A. Nelson, E. G. Wilkins, A. L. Pusic, and C. Sidey-Gibbons, "Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPIRED-001)," *The Breast*, vol. 60, pp. 111–122, Dec. 2021. Publisher: Elsevier.
41. Pedersen CF, Andersen M, Carreon LY, Eiskjær S. Applied Machine Learning for Spine Surgeons: Predicting Outcome for Patients Undergoing Treatment for Lumbar Disc Herniation Using PRO Data. *Global Spine Journal*. 2022;12:866–76.
42. A. Katakam, A. V. Karhade, A. Collins, D. Shin, C. Bragdon, A. F. Chen, C. M. Melnic, J. H. Schwab, and H. S. Bedair, "Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty," *Journal of Orthopaedic Research*, vol. 40, no. 4, pp. 808–815, 2022. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1002/jor.25125>.
43. C. B. Josephson, J. D. T. Engbers, T. T. Sajobi, S. Wahby, O. A. Lawal, M. R. Keezer, D. K. Nguyen, K. Malmgren, M. J. Atkinson, W. J. Hader, S. Macrodimitris, S. B. Patten, N. Pillay, R. Sharma, S. Singh, Y. Starreveld, and S. Wiebe, "Predicting postoperative epilepsy surgery satisfaction in adults using the 19-item Epilepsy Surgery Satisfaction Questionnaire and machine learning," *Epilepsia*, vol. 62, no. 9, pp. 2103–2112, 2021. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1111/epi.16992>.
44. P. N. Ramkumar, J. M. Karnuta, H. S. Haeberle, S. A. Rodeo, B. U. Nwachukwu, and R. J. Williams, "Effect of Preoperative Imaging and Patient Factors on Clinically Meaningful Outcomes and Quality of Life After Osteochondral Allograft Transplantation: A Machine Learning Analysis of Cartilage Defects of the Knee," *The American Journal of Sports Medicine*, vol. 49, pp. 2177–2186, 2021. Publisher: SAGE Publications Inc STM.
45. J. S. Munn, B. A. Lanting, S. J. MacDonald, L. E. Somerville, J. D. Marsh, D. M. Bryant, and B. M. Chesworth, "Logistic Regression and Machine Learning Models Cannot Discriminate Between Satisfied and Dissatisfied Total Knee Arthroplasty Patients," *The Journal of Arthroplasty*, vol. 37, pp. 267–273, 2022. Publisher: Elsevier.
46. Kober KM, Roy R, Dhruva A, Conley YP, Chan RJ, Cooper B, Olshen A, Miaskowski C. Prediction of evening fatigue severity in outpatients receiving chemotherapy: less may be more. *Fatigue : biomedicine, health & behavior*. 2021;9(1):14–32.
47. Durand WM, Daniels AH, Hamilton DK, Passias P, Kim HJ, Protosaltis T, LaFage V, Smith JS, Shaffrey C, Gupta M, Klineberg E, Schwab F, Burton D, Bess S, Ames C, Hart R. Artificial Intelligence Models Predict Operative Versus Nonoperative Management of Patients with Adult Spinal Deformity with 86% Accuracy. *World Neurosurgery*. 2020;141:e239–53.
48. C. J. Harrison, L. Geoghegan, C. J. Sidey-Gibbons, P. H. C. Stirling, J. E. McEachan, and J. N. Rodrigues, "Developing Machine Learning Algorithms to Support Patient-centered, Value-based Carpal Tunnel

- Decompression Surgery," *Plastic and Reconstructive Surgery Global Open*, vol. 10, p. e4279, Apr. 2022. V. Kumar, C. Roche, S. Overman, R. Simovitch, P.-H. Flurin, T. Wright, J. Zuckerman, H. Routman, and Teredesai, "What Is the Accuracy of Three Different Machine Learning Techniques to Predict Clinical Outcomes After Shoulder Arthroplasty?," *Clinical Orthopaedics and Related Research*, vol. 478, p. 2351, 2020.
49. Wshah S, Skalka C, Price M. Predicting Posttraumatic Stress Disorder Risk: A Machine Learning Approach", *JMIR Mental Health*, vol. 6, p. e13946, 2019. Y. Lu, E. Forlenza, R. R. Wilbur, O. Lavoie-Gagne, M. C. Fu, A. B. Yanke, B. J. Cole, N. Verma, and Forsythe, "Machine-learning model successfully predicts patients at risk for prolonged postoperative opioid use following elective knee arthroscopy. Knee surgery, sports traumatology, arthroscopy: official journal of the ESSKA. 2022;30:762–72.
50. Kunze KN, Karhade AV, Sadauskas AJ, Schwab JH, Levine BR. Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty. *J Arthroplasty*. 2020;35:2119–23.
51. N. L. Loos, L. Hoogendam, J. S. Souer, H. P. Slijper, E.-R. Andrinopoulou, M. W. Coppieters, R. W. Selles, and, the Hand-Wrist Study Group, "Machine Learning Can Be Used to Predict Function but Not Pain After Surgery for Thumb Carpometacarpal Osteoarthritis," *Clinical Orthopaedics and Related Research*, vol. 480, pp. 1271–1284, 2022.
52. Shipston-Sharman O, Popkirov S, Hansen CH, Stone J, Carson A. Prognosis in functional and recognised pathophysiological neurological disorders - a shared basis. *J Psychosom Res*. 2021;152: 110681.
53. B. K. Tan, G. Lu, M. J. Kwasny, W. D. Hsueh, S. Shintani-Smith, D. B. Conley, R. K. Chandra, R. C. Kern, and R. Leung, "Effect of symptom-based risk stratification on the costs of managing patients with chronic rhinosinusitis symptoms," *International Forum of Allergy & Rhinology*, vol. 3, no. 11, pp. 933–940, 2013. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1002/alar.21208>.
54. A. A. Rahman, M. I. Siraji, L. I. Khalid, F. Faisal, M. M. Nishat, A. Ahmed, and M. A. A. Mamun, "Perceived Stress Analysis of Undergraduate Students during COVID-19: A Machine Learning Approach," in 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON), pp. 1129–1134, 2022. ISSN: 2158–8481.
55. E. S. Lee, "Exploring the Performance of Stacking Classifier to Predict Depression Among the Elderly," in 2017 IEEE International Conference on Healthcare Informatics (ICHI), pp. 13–20, 2017.
56. Nowinka Z, Alagha MA, Mahmoud K, Jones GG. Predicting Depression in Patients With Knee Osteoarthritis Using Machine Learning: Model Development and Validation Study. *JMIR formative research*. 2022;6: e36130.
57. Arkin FS, Aras G, Dogu E. Comparison of Artificial Neural Networks and Logistic Regression for 30-days Survival Prediction of Cancer Patients. *Acta informatica medica: AIM: journal of the Society for Medical Informatics of Bosnia & Herzegovina: casopis Društva za medicinsku informatiku BiH*. 2020;28:108–13.
58. L. Hoogendam, J. A. C. Bakx, J. S. Souer, H. P. Slijper, E.-R. Andrinopoulou, R. W. Selles, and Hand Wrist Study Group, "Predicting Clinically Relevant Patient-Reported Symptom Improvement After Carpal Tunnel Release: A Machine Learning Approach," *Neurosurgery*, vol. 90, pp. 106–113, 2022.
59. Camp EJ, Quon RJ, Sajatovic M, Briggs F, Brownrigg B, Janevic MR, Meisenhelter S, Steimel SA, Testorf ME, Kiriakopoulos E, Mazanec MT, Fraser RT, Johnson EK, Jobst BC. "Supervised machine learning to predict reduced depression severity in people with epilepsy through epilepsy self-management intervention", *Epilepsy & Behavior*, vol. 127. Elsevier: Publisher; 2022.
60. Wardenaar KJ, Riese H, Giltay EJ, Eikelenboom M, van Hemert AJ, Beekman AF, Penninx BWJH, Schoevers RA. Common and specific determinants of 9-year depression and anxiety course-trajectories: A machine-learning investigation in the Netherlands Study of Depression and Anxiety (NESDA). *J Affect Disord*. 2021;293:295–304.
61. J.-a. Sim, Y. A. Kim, J. H. Kim, J. M. Lee, M. S. Kim, Y. M. Shim, J. I. Zo, and Y. H. Yun, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Scientific Reports*, vol. 10, p. 10693, 2020. Number: 1 Publisher: Nature Publishing Group.
62. K. N. Kunze, E. M. Polce, B. U. Nwachukwu, J. Chahla, and S. J. Nho, "Development and Internal Validation of Supervised Machine Learning Algorithms for Predicting Clinically Significant Functional Improvement in a Mixed Population of Primary Hip Arthroscopy," *Arthroscopy*, vol. 37, pp. 1488–1497, 2021. Publisher: Elsevier.
63. Smith DL, Held P. Moving toward precision PTSD treatment: predicting veterans' intensive PTSD treatment response using continuously updating machine learning models. *Psychol Med*. 2023;53:5500–9.
64. Bone C, Simmonds-Buckley M, Thwaites R, Sandford D, Merzhvynska M, Rubel J, Deisenhofer A-K, Lutz W, Delgadillo J. Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*. 2021;3:e231–40.
65. O'Driscoll C, Buckman JEJ, Fried EI, Saunders R, Cohen ZD, Ambler G, DeRubeis RJ, Gilbody S, Hollon SD, Kendrick T, Kessler D, Lewis G, Watkins E, Wiles N, Pilling S. The importance of transdiagnostic symptom level assessment to understanding prognosis for depressed adults: analysis of data from six randomised control trials. *BMC Med*. 2021;19:109.
66. Shi H-Y, Tsai J-T, Chen Y-M, Culbertson R, Chang H-T, Hou M-F. Predicting two-year quality of life after breast cancer surgery using artificial neural network and linear regression models. *Breast Cancer Res Treat*. 2012;135:221–9.
67. Pua Y-H, Kang H, Thumboo J, Clark RA, Chew ES-X, Poon CL-L, Chong H-C, Yeo S-J. Machine learning methods are comparable to logistic regression techniques in predicting severe walking limitation following total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2020;28:3207–16.
68. Langenberger B. Who will stay a little longer? Predicting length of stay in hip and knee arthroplasty patients using machine learning. *Intelligence-Based Medicine*. 2023;8: 100111.
69. Hasannejadasl H, Osong B, Bermejo I, van der Poel H, Vanneste B, van Roermund J, Aben K, Zhang Z, Kiemeny L, Van Oort I, Verwey R, Hochstenbach L, Bloemen E, Dekker A, Fijten RRR. A comparison of machine learning models for predicting urinary incontinence in men with localized prostate cancer. *Front Oncol*. 2023;13:1168219.
70. Xu C, Pfof A, Mehrara BJ, Yin P, Nelson JA, Pusic AL, Sidey-Gibbons C. Enhanced Surgical Decision-Making Tools in Breast Cancer: Predicting 2-Year Postoperative Physical, Sexual, and Psychosocial Well-Being following Mastectomy and Breast Reconstruction (INSPIRED 004). *Ann Surg Oncol*. 2023;30:7046–59.
71. Tian J, Yan J, Han G, Du Y, Hu X, He Z, Han Q, Zhang Y. Machine learning prognosis model based on patient-reported outcomes for chronic heart failure patients after discharge. *Health Qual Life Outcomes*. 2023;21:31.
72. C. Park, P. V. Mummaneni, O. N. Gottfried, C. I. Shaffrey, A. J. Tang, E. F. Bisson, A. L. Asher, D. Coric, E. A. Potts, K. T. Foley, M. Y. Wang, K.-M. Fu, M. S. Virk, J. J. Knightly, S. Meyer, P. Park, Upadhyaya, M. E. Shaffrey, A. L. Buchholz, L. M. Tumial'an, J. D. Turner, B. A. Sherrod, N. Agarwal, D. Chou, R. W. Haid, M. Bydon, and A. K. Chan, "Which supervised machine learning algorithm can best predict achievement of minimum clinically important difference in neck pain after surgery in patients with cervical myelopathy? A QOD study," *Neurosurgical Focus*, vol. 54, p. E5, 2023. Publisher: American Association of Neurological Surgeons Section: Neurosurgical Focus.
73. Langenberger B, Schrednitzki D, Halder AM, Busse R, Pross CM. Predicting whether patients will achieve minimal clinically important differences following hip or knee arthroplasty. *Bone & Joint Research*. 2023;12:512–21.
74. Curtis JR, Su Y, Black S, Xu S, Langholf W, Bingham CO, Kafka S, Xie F. Machine Learning Applied to Patient-Reported Outcomes to Classify Physician-Derived Measures of Rheumatoid Arthritis Disease Activity. *ACR open rheumatology*. 2022;4:995–1003.
75. Iivanainen S, Ekstrom J, Virtanen H, Katja VV, Koivunen JP. Electronic patient-reported outcomes and machine learning in predicting immune-related adverse events of immune checkpoint inhibitor therapies. *BMC Med Inform Decis Mak*. 2021;21:205.
76. Pfof A, Mehrara BJ, Nelson JA, Wilkins EG, Pusic AL, Sidey-Gibbons C. Towards Patient-centered Decision-making in Breast Cancer Surgery: Machine Learning to Predict Individual Patient-reported Outcomes at 1-year Follow-up. *Ann Surg*. 2023;277: e144.
77. C. W. Noel, R. Sutradhar, L. Gotlib Conn, D. Forner, W. C. Chan, R. Fu, J. Hallet, N. G. Coburn, and A. Eskander, "Development and Validation of

- a Machine Learning Algorithm Predicting Emergency Department Use and Unplanned Hospitalization in Patients With Head and Neck Cancer," *JAMA otolaryngology— head & neck surgery*, vol. 148, pp. 764–772, 2022.
78. Chmiel FP, Burns DK, Pickering JB, Blythin A, Wilkinson TM, Boniface MJ. Prediction of Chronic Obstructive Pulmonary Disease Exacerbation Events by Using Patient Self-reported Data in a Digital Health App: Statistical Evaluation and Machine Learning Approach. *JMIR Med Inform*. 2022;10: e26499.
79. H. N. Ziobrowski, C. J. Kennedy, B. Ustun, S. L. House, F. L. Beaudoin, X. An, D. Zeng, K. A. Bollen, M. Petukhova, N. A. Sampson, V. Puac-Polanco, S. Lee, K. C. Koenen, K. J. Ressler, S. A. McLean, R. C. Kessler, AURORA Consortium, J. S. Stevens, T. C. Neylan, G. D. Clifford, T. Jovanovic, S. D. Linnstaedt, L. T. Germin, S. L. Rauch, J. P. Haran, A. B. Storrow, C. Lewandowski, P. I. Musey, P. L. Hendry, S. Sheikh, C. W. Jones, B. E. Patches, M. S. Lyons, V. P. Murty, M. E. McGrath, J. L. Pascual, M. J. Seamon, E. M. Datner, A. M. Chang, C. Pearson, D. A. Peak, G. Jambaulikar, R. C. Merchant, R. M. Domeier, N. K. Rathlev, B. J. O'Neil, P. Sergot, L. D. Sanchez, S. E. Bruce, R. H. Pietrzak, J. Joormann, D. M. Barch, D. A. Pizzagalli, J. F. Sheridan, S. E. Harte, J. M. Elliott, and S. J. H. van Rooij, "Development and Validation of a Model to Predict Posttraumatic Stress Disorder and Major Depression After a Motor Vehicle Collision," *JAMA psychiatry*, vol. 78, pp. 1228–1237, 2021.
80. P. Annareddy, M. F. Hossain, T. Kissane, W. Frydrychowicz, P. Nitu, J. Coelho, N. Johnson, P. Madiraju, Z. Franco, K. Hooyer, N. Jain, M. Flower, and S. Ahamed, "Predicting PTSD Severity in Veterans from Self-reports for Early Intervention: A Machine Learning Approach," in 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), pp. 201–208, 2020.
81. M. Zhong, V. Van Zoest, A. M. Bilal, F. Papadopoulos, and G. Castellano, "Unimodal vs. Multimodal Prediction of Antenatal Depression from Smartphone-based Survey Data in a Longitudinal Study," in Proceedings of the 2022 International Conference on Multimodal Interaction, ICMI '22, (New York, NY, USA), pp. 455–467, Association for Computing Machinery, 2022.
82. C. J. Sidey-Gibbons, C. Sun, A. Schneider, S.-C. Lu, K. Lu, A. Wright, and L. Meyer, "Predicting 180day mortality for women with ovarian cancer using machine learning and patient-reported outcome data," *Scientific Reports*, vol. 12, p. 21269, 2022. Number: 1 Publisher: Nature Publishing Group.
83. L. S. Canas, C. H. Sudre, J. C. Pujol, L. Polidori, B. Murray, E. Molteni, M. S. Graham, K. Klaser, M. Antonelli, S. Berry, R. Davies, L. H. Nguyen, D. A. Drew, J. Wolf, A. T. Chan, T. Spector, C. J. Steves, S. Ourselin, and M. Modat, "Early detection of COVID-19 in the UK using self-reported symptoms: a large-scale, prospective, epidemiological surveillance study," *The Lancet Digital Health*, vol. 3, pp. e587– e598, 2021. Publisher: Elsevier.
84. R. Sutradhar and L. Barbera, "Comparing an Artificial Neural Network to Logistic Regression for Predicting ED Visit Risk Among Patients With Cancer: A Population-Based Cohort Study," *Journal of Pain and Symptom Management*, vol. 60, pp. 1–9, 2020. Publisher: Elsevier.
85. Goldstein N, Eisenkraft A, Arguello CJ, Yang GJ, Sand E, Ishay AB, Merin R, Fons M, Littman R, Nachman D, Gepner Y. Exploring Early Pre-Symptomatic Detection of Influenza Using Continuous Monitoring of Advanced Physiological Parameters during a Randomized Controlled Trial. *J Clin Med*. 2021;10:5202.
86. Rahman QA, Janmohamed T, Clarke H, Ritvo P, Heffernan J, Katz J. Interpretability and Class Imbalance in Prediction Models for Pain Volatility in Manage My Pain App Users: Analysis Using Feature Selection and Majority Voting Methods. *JMIR Med Inform*. 2019;7: e15601.
87. Verma D, Jansen D, Bach K, Poel M, Mork PJ, d'Hollosy WON. Exploratory application of machine learning methods on patient reported data in the development of supervised models for predicting outcomes. *BMC Med Inform Decis Mak*. 2022;22:227.
88. Crowson MG, Dixon P, Mahmood R, Lee JW, Shipp D, Le T, Lin V, Chen J, Chan TCY. Predicting Postoperative Cochlear Implant Performance Using Supervised Machine Learning. *Otol Neurotol*. 2020;41: e1013.
89. Miella F, Famiglini L, Banfi G, Cabitza F. Application of Machine Learning to Improve Appropriateness of Treatment in an Orthopaedic Setting of Personalized Medicine. *Journal of Personalized Medicine*. 2022;12:1706.
90. E. M. Polce, K. N. Kunze, M. C. Fu, G. E. Garrigues, B. Forsythe, G. P. Nicholson, B. J. Cole, and N. N. Verma, "Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty," *Journal of Shoulder and Elbow Surgery*, vol. 30, pp. e290–e299, 2021. Publisher: Elsevier.
91. P. Thanathamthee, "Boosting with feature selection technique for screening and predicting adolescents depression," in 2014 Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), pp. 23–27, 2014.
92. J. A. A. Mendoza, G. A. Solano, M. J. Pontiveros, J. DI Caro, P. M. D. Gomez, C. G. Manuel, P. J. B. Rosell-Ubial, and M. Tee, "A Machine Learning Approach in Evaluating Symptom Screening in Predicting COVID-19," in 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIC), pp. 188–193, 2022.
93. Z. Xu, Y. Wang, and B. Jiang, "Research on Prisoner Psychological Symptoms Quick Screening Model Based on Ensemble Learning," in Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences, ISAIMS '22, (New York, NY, USA), pp. 563–567, Association for Computing Machinery, Dec. 2022.
94. Fu MR, Wang Y, Li C, Qiu Z, Axelrod D, Guth AA, Scagliola J, Conley Y, Aouizerat BE, Qiu JM, Yu G, Cleave JHV, Haber J, Cheung YK. "Machine learning for detection of lymphedema among breast cancer survivors", *mHealth*, vol. 4. AME publishing company: Publisher; 2018.
95. X. Pan, R. Levin-Epstein, J. Huang, D. Ruan, C. R. King, A. U. Kishan, M. L. Steinberg, and X. S. Qi, "Dosimetric predictors of patient-reported toxicity after prostate stereotactic body radiotherapy: Analysis of full range of the dose–volume histogram using ensemble machine learning," *Radiotherapy and Oncology*, vol. 148, pp. 181–188, 2020. Publisher: Elsevier.
96. Agochukwu-Mmonu N, Murali A, Wittmann D, Denton B, Dunn RL, Montie J, Peabody J, Miller D, Singh K. Development and Validation of Dynamic Multivariate Prediction Models of Sexual Function Recovery in Patients with Prostate Cancer Undergoing Radical Prostatectomy: Results from the MUSIC Statewide Collaborative. *European Urology Open Science*. 2022;40:1–8.
97. A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorgieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett, "Cross-trial prediction of treatment outcome in depression: a machine learning approach," *The Lancet Psychiatry*, vol. 3, pp. 243–250, 2016. Publisher: Elsevier.
98. W. Oude Nijeweme-d'Hollosy, L. van Velsen, M. Poel, C. G. M. Groothuis-Oudshoorn, R. Soer, and H. Hermens, "Evaluation of three machine learning models for self-referral decision support on low back pain in primary care," *International Journal of Medical Informatics*, vol. 110, pp. 31–41, 2018.
99. A. Goldstein and S. Cohen, "Self-report symptom-based endometriosis prediction using machine learning," *Scientific Reports*, vol. 13, p. 5499, 2023. Number: 1 Publisher: Nature Publishing Group.
100. S. Iivanainen, J. Ekstr om, H. Virtanen, V. V. Kataja, and J. P. Koivunen, "Predicting Objective Response Rate (ORR) in Immune Checkpoint Inhibitor (ICI) Therapies with Machine Learning (ML) by Combining Clinical and Patient-Reported Data," *Applied Sciences*, vol. 12, p. 1563, 2022. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
101. G. Luo, B.L. Stone, B. Fassl, et al. "Predicting asthma control deterioration in children." *BMC Med Inform Decis Mak* 15, 84, (2015)
102. Verma D, Bach K, Mork PJ. External validation of prediction models for patient-reported outcome measurements collected using the selfBACK mobile app. *Int J Med Informatics*. 2023;170: 104936.
103. Haeberle HS, Ramkumar PN, Karnuta JM, Sullivan R, Sink EL, Kelly BT, Ranawat AS, Nwachukwu BU. Predicting the risk of subsequent hip surgery before primary hip arthroscopy for femoroacetabular impingement syndrome: a machine learning analysis of preoperative risk factors in hip preservation. *Am J Sports Med*. 2021;49:2668–76.
104. Sun R, Tomkins-Lane C, Muaremi A, Kuwabara A, Smuck M. Physical activity thresholds for predicting longitudinal gait decline in adults with knee osteoarthritis. *Osteoarthritis Cartilage*. 2021;29:965–72.
105. K. Schultebrucks, M. Qian, D. Abu-Amara, K. Dean, E. Laska, C. Siegel, A. Gautam, G. Guffanti, R. Hammamieh, B. Misganaw, S. H. Mellon, O. M. Wolkowitz, E. M. Blessing, A. Etkin, K. J. Ressler, F. J. Doyle, M. Jett, and C. R. Marmar, "Pre-deployment risk factors for PTSD in active-duty personnel deployed to Afghanistan: a machine-learning approach

- for analyzing multivariate predictors," *Molecular Psychiatry*, vol. 26, pp. 5011–5022, 2021. Number: 9 Publisher: Nature Publishing Group.
106. Tschuggnall M, Grote V, Pirchl M, Holzner B, Rumpold G, Fischer MJ. Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Informatics in Medicine Unlocked*. 2021;24: 100598.
107. M. H. Sandham, E. A. Hedgecock, R. J. Siegert, A. Narayanan, M. B. Hocaoglu, and I. J. Higginson, "Intelligent Palliative Care Based on Patient-Reported Outcome Measures," *Journal of Pain and Symptom Management*, vol. 63, pp. 747–757, 2022. Publisher: Elsevier.
108. H. Pappot, B. P. Björnsson, O. Krause, C. Bæksted, P. E. Bidstrup, S. O. Dalton, C. Johansen, A. Knoop, I. Vogelijs, and C. Hollander-Mieritz, "Machine learning applied in patient-reported outcome research—exploring symptoms in adjuvant treatment of breast cancer," *Breast Cancer*, vol. 31, pp. 148–153, 2024.
109. Y.-C. Hsu, J.-D. Wang, P.-H. Huang, Y.-W. Chien, C.-J. Chiu, and C.-Y. Lin, "Integrating domain knowledge with machine learning to detect obstructive sleep apnea: Snore as a significant bio-feature," *Journal of Sleep Research*, vol. 31, no. 2, p. e13487, 2022. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1111/jsr.13487>.
110. I. Miranda, G. Cardoso, M. Pahar, G. Oliveira, and T. Niesler, "Machine Learning Prediction of Hospitalization due to COVID-19 based on Self-Reported Symptoms: A Study for Brazil," in 2021 IEEEEMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–5, 2021. ISSN: 2641–3604.
111. Long MJ, Papi E, Duffell LD, McGregor AH. Predicting knee osteoarthritis risk in injured populations. *Clin Biomech*. 2017;47:87–95.
112. Xuyi W, Seow H, Sutradhar R. Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer. *Cancer Med*. 2020;10:989–98.
113. Shahzad MN, Suleman M, Ahmed MA, Riaz A, Fatima K. Identifying the Symptom Severity in Obsessive-Compulsive Disorder for Classification and Prediction: An Artificial Neural Network Approach. *Behav Neurol*. 2020;2020:2678718.
114. A. Bugajski, A. Lengerich, R. Koerner, and L. Szalacha, "Utilizing an Artificial Neural Network to Predict Self-Management in Patients With Chronic Obstructive Pulmonary Disease: An Exploratory Analysis," *Journal of Nursing Scholarship*, vol. 53, no. 1, pp. 16–24, 2021. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1111/jnu.12618>.
115. L. van der Stap, M. F. van Haften, E. F. van Marrewijk, A. H. de Heij, P. L. Jansen, J. M. N. Burgers, M. S. Sieswerda, R. K. Los, A. K. L. Reyniers, and Y. M. van der Linden, "The feasibility of a Bayesian network model to assess the probability of simultaneous symptoms in patients with advanced cancer," *Scientific Reports*, vol. 12, p. 22295, 2022. Number: 1 Publisher: Nature Publishing Group.
116. T. Strating, L. Shafiee Hanjani, I. Tornvall, R. Hubbard, and I. A. Scott, "Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models," *BMJ Health & Care Informatics*, vol. 30, p. e100767, 2023.
117. S. Cruz Rivera, X. Liu, S. E. Hughes, H. Dunster, E. Manna, A. K. Denniston, and M. J. Calvert, "Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies," *The Lancet Digital Health*, vol. 5, pp. e168–e173, 2023.
118. Whitebird RR, Solberg LI, Ziegenfuss JY, Norton CK, Chrenka EA, Swiontkowski M, Reams M, Grossman ES. What Do Orthopaedists Believe is Needed for Incorporating Patient-reported Outcome Measures into Clinical Care? A Qualitative Study. *Clin Orthop Relat Res*. 2022;480:680–7.
119. M. R. Davies, J. E. J. Buckman, B. N. Adey, C. Armour, J. R. Bradley, S. C. B. Curzons, H. L. Davies, K. A. S. Davis, K. A. Goldsmith, C. R. Hirsch, M. Hotopf, C. Hu, bel, I. R. Jones, G. Kalsi, G. Krebs, Y. Lin, I. Marsh, M. McAtarsney-Kovacs, A. M. McIntosh, J. Mundy, D. Monssen, A. J. Peel, H. C. Rogers, M. Skelton, D. J. Smith, A. ter Kuile, K. N. Thompson, D. Veale, J. T. R. Walters, R. Zahn, G. Breen, and T. C. Eley, "Comparison of symptom-based versus self-reported diagnostic measures of anxiety and depression disorders in the GLAD and COPING cohorts," *Journal of Anxiety Disorders*, vol. 85, p. 102491, 2022.
120. S. Schmeelk, A. Davis, Q. Li, C. Shippey, M. Utah, A. Myers, M. R. Turchioe, and R. M. Creber, "Monitoring Symptoms of COVID-19: Review of Mobile Apps," *JMIR mHealth and uHealth*, vol. 10, p. e36065, June 2022. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.
121. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial Intelligence Applications in Health Care Practice: Scoping Review. *J Med Internet Res*. 2022;24: e40238.
122. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59:1087–91.
123. D. M. P. Murti, U. Pujiarto, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," in 2019 5th International Conference on Science in Information Technology (ICSITech), pp. 83–88, 2019.
124. Y. Liu, Y. Li, and D. Xie, "Implications of imbalanced datasets for empirical ROC-AUC estimation in binary classification tasks," *Journal of Statistical Computation and Simulation*, vol. 94, pp. 183–203, 2024. Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/00949655.2023.2238235>.
125. W. Zuzanna, Dimitrova Vania, Warrington Lorraine, Velikova Galina, and Absolom Kate, "Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From PatientReported Outcome Measures," *JCO Clinical Cancer Informatics (In press)*, 2024.
126. D. Jha, A. Rauniyar, A. Srivastava, D. H. Hagos, N. K. Tomar, V. Sharma, E. Keles, Z. Zhang, U. Demir, A. Topcu, A. Yazidi, J. E. H. "aakeg"ard, and U. Bagci, "Ensuring Trustworthy Medical Artificial Intelligence through Ethical and Philosophical Principles," 2023. [arXiv:2304.11530](https://arxiv.org/abs/2304.11530) [cs].
127. Nisar D-E-M, Amin R, Shah N-U-H, Ghamdi MAA, Almotiri SH, Alruily M. Healthcare Techniques Through Deep Learning: Issues, Challenges and Opportunities. *IEEE Access*. 2021;9:98523–41.
128. Morid MA, Sheng ORL, Dunbar J. Time Series Prediction Using Deep Learning Methods in Healthcare. *ACM Trans Manag Inf Syst*. 2023;14:1–29.
129. I. R. Ward, L. Wang, J. lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable Artificial Intelligence for Pharmacovigilance: What Features Are Important When Predicting Adverse Outcomes?," *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106415, 2021. [arXiv:2112.13210](https://arxiv.org/abs/2112.13210) [cs, q-bio].
130. Kelly BS, Kirwan A, Quinn MS, Kelly AM, Mathur P, Lawlor A, Killeen RP. The ethical matrix as a method for involving people living with disease and the wider public (PPI) in near-term artificial intelligence research. *Radiography*. 2023;29:S103–11.
131. Y. Chen, A. A. Hosin, M. J. George, F. W. Asselbergs, and A. D. Shah, "Digital technology and patient and public involvement (PPI) in routine care and clinical research—A pilot study," *PLOS ONE*, vol. 18, p. e0278260, 2023. Publisher: Public Library of Science.
132. W. Xuyi, H. Seow, and R. Sutradhar, "Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer," *Cancer Medicine*, vol. 10, no. 3, pp. 989–998, 2021. eprint: <https://onlinelibrary.wiley.com/doi/pdf/https://doi.org/10.1002/cam4.3685>.
133. L.-m. Wang, J.-x. Chen, Y. Pei, X. Zhao, H.-t. Cui, and H.-z. Cui, "Feature Selection and Prediction of Sub-health State Using SVM-RFE," in 2010 International Conference on Artificial Intelligence and Computational Intelligence, vol. 3, pp. 199–202, 2010.

Publisher's Note

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

8.2 Paper 2

Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From Patient-Reported Outcome Measures

Zuzanna Wójcik, MSc¹ ; Vania Dimitrova, PhD² ; Lorraine Warrington, PhD³ ; Galina Velikova, BMBS(MD), PhD^{3,4} ; and Kate Absolom, PhD^{3,5} 

DOI <https://doi.org/10.1200/CCI.23.00264>

ABSTRACT

ACCOMPANYING CONTENT

Appendix

Accepted March 1, 2024

Published April 26, 2024

JCO Clin Cancer Inform

8:e2300264

© 2024 by American Society of
Clinical Oncology

PURPOSE Adverse effects of chemotherapy often require hospital admissions or treatment management. Identifying factors contributing to unplanned hospital utilization may improve health care quality and patients' well-being. This study aimed to assess if patient-reported outcome measures (PROMs) improve performance of machine learning (ML) models predicting hospital admissions, triage events (contacting helpline or attending hospital), and changes to chemotherapy.

MATERIALS AND METHODS Clinical trial data were used and contained responses to three PROMs (European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire [QLQ-C30], EuroQol Five-Dimensional Visual Analogue Scale [EQ-5D], and Functional Assessment of Cancer Therapy-General [FACT-G]) and clinical information on 508 participants undergoing chemotherapy. Six feature sets (with following variables: [1] all available; [2] clinical; [3] PROMs; [4] clinical and QLQ-C30; [5] clinical and EQ-5D; [6] clinical and FACT-G) were applied in six ML models (logistic regression [LR], decision tree, adaptive boosting, random forest [RF], support vector machines [SVMs], and neural network) to predict admissions, triage events, and chemotherapy changes.

RESULTS The comprehensive analysis of predictive performances of the six ML models for each feature set in three different methods for handling class imbalance indicated that PROMs improved predictions of all outcomes. RF and SVMs had the highest performance for predicting admissions and changes to chemotherapy in balanced data sets, and LR in imbalanced data set. Balancing data led to the best performance compared with imbalanced data set or data set with balanced train set only.

CONCLUSION These results endorsed the view that ML can be applied on PROM data to predict hospital utilization and chemotherapy management. If further explored, this study may contribute to health care planning and treatment personalization. Rigorous comparison of model performance affected by different imbalanced data handling methods shows best practice in ML research.

Licensed under the Creative
Commons Attribution 4.0 License

INTRODUCTION

Cancer treatment side effects frequently negatively affect patients' health and often cause emergency hospitalization.^{1,2} Unplanned health care utilization can be detrimental for patients' physical and emotional well-being and can reduce health care quality through burdening health care systems.³ Early identification of factors contributing to acute hospital presentations can support planning for emergency admissions, increase the quality of care, and reduce health care costs.^{2,4} Predicting the risk of chemotherapy-related hospital

utilization could also help personalizing cancer treatment decisions.^{5,6}

Machine learning (ML) adoption in medicine can aid clinical decisions, improving health care quality.⁷ ML methods have been applied to predict health outcomes, including post-surgery complications,⁸ stroke rehabilitation success,⁹ epilepsy,¹⁰ or mortality.¹¹ ML models can also be successful in predicting hospital utilization. For instance, binary classifiers were used to robustly predict hospital admissions on the basis of emergency department triage information

CONTEXT

Key Objective

To assess if patient-reported outcome measures (PROMs) improve performance of machine learning (ML) models predicting hospital admissions, triage events, and changes to chemotherapy.

Knowledge Generated

PROMs improve ML models predicting unplanned hospital utilization and chemotherapy management. ML methods provide good performance predicting changes to chemotherapy.

Relevance (J.L. Warner)

This study is one of the first published by this journal to show an improvement in ML model performance from including patient-reported outcomes. The authors plan to involve patients and clinicians to assess their attitudes to ML-based prediction in their future work.*

*Relevance section written by JCO Clinical Cancer Informatics Editor-in-Chief Jeremy L. Warner, MD, MS, FAMIA, FASCO.

and patients' medical history.¹² Furthermore, ML algorithms were applied to electronic health records (EHR) to predict chemotherapy-related hospital admissions.⁵

However, these models did not include any information gathered from patients about their own health and well-being. Therefore, current AI models process the clinical information well, without consideration of patients' perspective on their health.

Patient-reported outcome measures (PROMs) are questionnaires that measure patients' perception on their own health status,¹³ including disease-related symptoms, side effects of treatments, quality of life, and impact on functioning. PROMs are increasingly incorporated in routine clinical care and can be used as predictors in ML methods foreseeing health outcomes,^{14,15} for example, identifying patients at risk of experiencing undesirable clinical outcomes.¹⁶ ML algorithms trained on patient-reported and clinical data accurately predicted financial toxicity in patients with early breast cancer.¹⁷ Furthermore, PROMs enhanced ML performance predicting 5-year cancer survival, when added to clinical and sociodemographic variables.¹⁸ Nevertheless, the benefits of inclusion of PROMs as predictors are inconsistent, as some studies did not find PROMs to have as meaningful impact on model performance as objective measures.^{19,20}

The variability in effectiveness of PROMs in predicting patient outcomes may be caused by inconsistent performance metrics and conclusions drawn from data affected by inappropriate preprocessing methods, such as balancing data sets before creating training and testing sets, which often introduces bias.²¹ The lack of methodologic agreement and guidance in the literature indicates the need for comparison of frequently used methods. The predictive value of PROMs is

also not explored in detail because of the variety of PROMs currently used.¹⁵ Therefore, this paper aims to address five research questions:

1. Do PROMs add predictive value to ML models?
2. Which PROMs are the most useful in predictions?
3. Which ML models have the best performance?
4. Did preprocessing method for handling class imbalance affect model performance?
5. Which features were the most important for prediction?

MATERIALS AND METHODS

Data Set

Data from 508 patients initiating systemic treatment for colorectal, breast, or gynecologic cancers at Leeds Cancer Centre (United Kingdom), collected in an eRAPID clinical trial between January 22, 2015, and June 11, 2018,²² were used in this study. The data set contained 35 variables. Eight variables were clinical or demographic, collected from EHR. They included age at study entry, sex (male/female), number of days on study from the start of chemotherapy, study arm, disease site (breast/gynecologic/colorectal), previous chemotherapy (yes/no), information if the disease was metastatic or nonmetastatic, and the number of comorbidities (from the list: cardiovascular, respiratory, gastrointestinal, stomach/intestine, endocrine, renal, neurologic, rheumatologic, previous malignancy, and substance abuse). Twenty-four variables were from PROMs completed by participants on paper at the time of study entry. Fifteen of these PROMs were from European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire (QLQ-C30)²³ with 30 items, containing information about participants' physical symptoms, perception on their physical function, emotional and social function, and overall

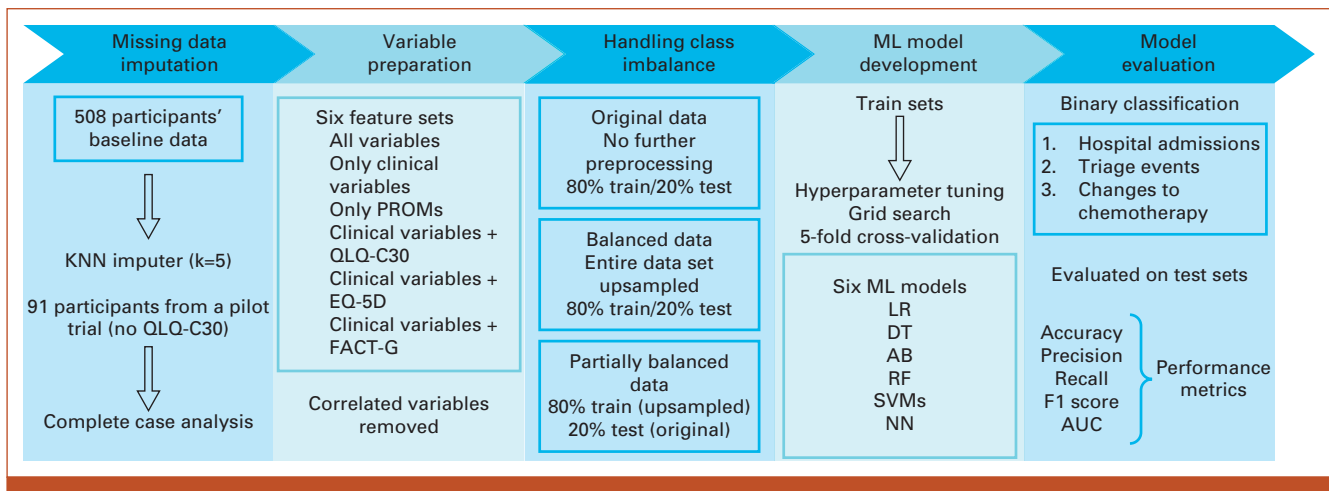


FIG 1. Flow diagram illustrating the methodology of the study. AB, adaptive boosting; DT, decision tree; EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; KNN, k-nearest neighbors; LR, logistic regression; ML, machine learning; NN, neural network; PROMs, patient-reported outcome measures; QLQ-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest; SVMs, support vector machines.

health and quality of life. Another five PROM variables were from Five-Dimensional Visual Analogue Scale (EQ-5D),²⁴ including self-reported data on mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Four remaining PROM variables were aggregated scores of physical, social, emotional, and functional well-being from Functional Assessment of Cancer Therapy-General (FACT-G) 28 items.²⁵ Three target variables were the number of hospital admissions, triage events (patients contacting emergency helpline or attending oncology admission unit), and changes to chemotherapy during the 18-week clinical trial. This information was extracted from EHR. The variables were selected because of their availability from the eRAPID clinical trial²² and the consultation with clinicians regarding their relevance.

Variable Preparation

The overview of the methods is presented in [Figure 1](#). Target features were transformed to binary variables with class 0 (no event) or 1 (at least one event) to enable binary classification.^{26,27} To allow in-depth exploration of all PROM effects on the model performance in general, and when individual questionnaires are separately added to clinical data, six different feature sets were created with following variables:

1. Only clinical
2. All available
3. Only PROMs
4. Clinical + QLQ-C30
5. Clinical + EQ-5D
6. Clinical + FACT-G

Continuous variables were scaled to unit variance to improve computational performance of ML.²⁸ To prevent algorithms from receiving repeated information,²⁹ correlated variables were removed from each feature set (leaving one), so that no

Pearson coefficient higher than 0.6 was left.³⁰ The list of variables in each feature set is presented in [Appendix Tables A1 and A2](#), including differences between classes.

Missing Data Imputation

All patients completed QLQ-C30, EQ-5D, and FACT-G at the clinical trial baseline. However, for 91 participants whose data were taken from the pilot study of the trial, only two subscales of QLQ-C30 were included, so patients from this phase did not have full QLQ-C30 data. The records from these participants were removed from affected feature sets (all variables, only PROMs, and clinical + QLQ-C30 variables). Using complete case analysis (CC) is justified under the missing completely at random assumption. Pilot trial ensures random selection of participants, so CC method is unlikely to bias results.³¹ Any further cases of missing values were infrequent and likely resulted from participants omitting questions, which is a common issue in PROM data.³² They were imputed using K-nearest neighbors algorithm ($k = 5$), being a common imputation method in relevant studies.^{18,26,33,34}

Handling Class Imbalance

To mitigate potential bias of class imbalance,²¹ synthetic participants in minority class can be created to match the number of participants in the majority class (oversampling). In previous studies, it was performed on the entire data set^{18,35} or training set only.^{36,37} ML can also be trained on original data and evaluated using multiple performance metrics.³³ Since there is no consistency in data preprocessing methods, the model performances in these scenarios were compared to discover bias in the results. Therefore, three data sets were created from each of the six feature sets for all target variables.

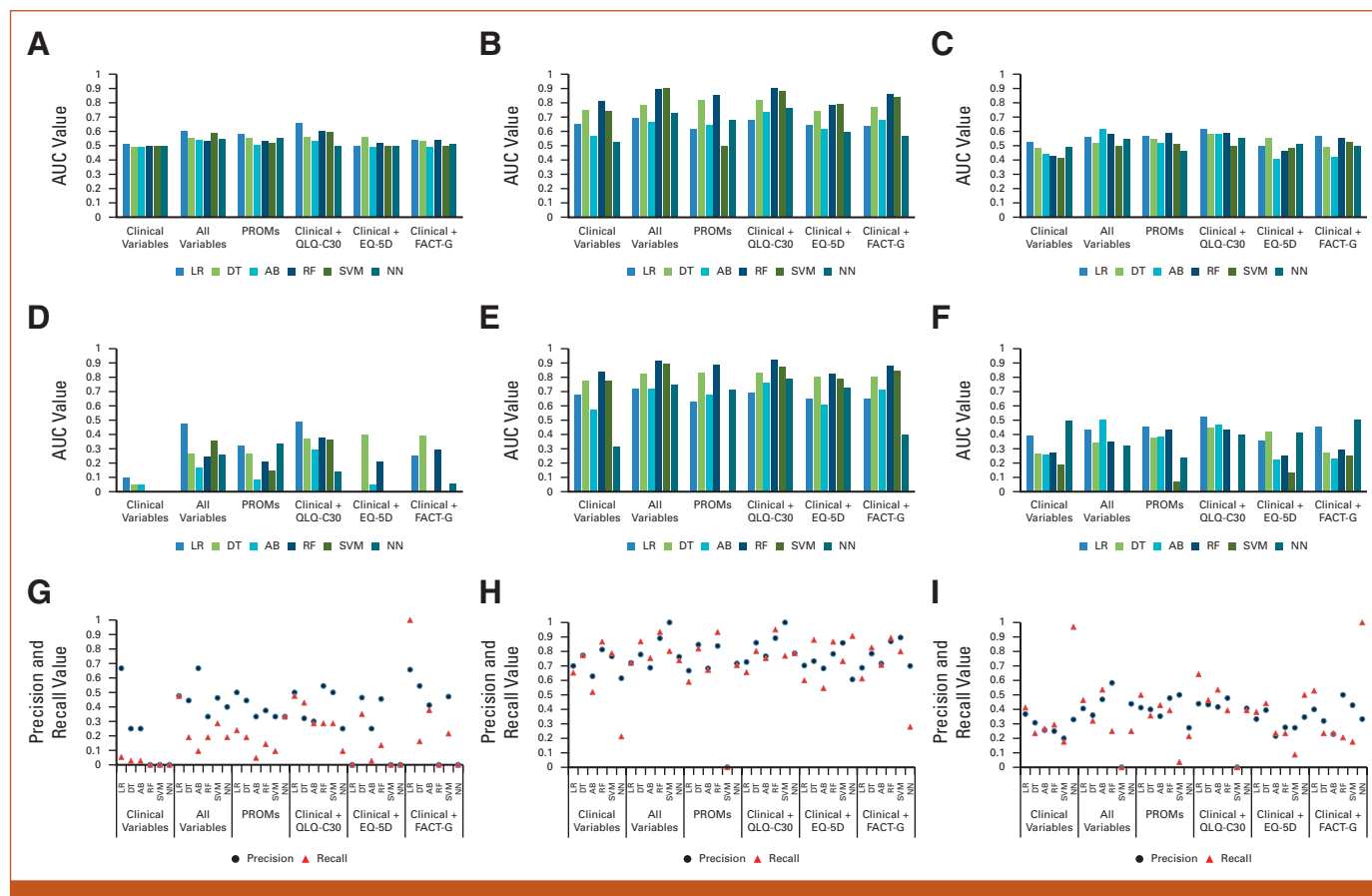


FIG 2. Predicting admissions. AUC values for (A) original, (B) balanced, and (C) partially balanced data sets obtained from all models in all feature sets; F1 scores for (D) original, (E) balanced, and (F) partially balanced data sets obtained from all models in all feature sets; precision and recall distributions for (G) original, (H) balanced, and (I) partially balanced data sets obtained from all models in all feature set. AB, adaptive boosting; DT, decision tree; EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; LR, logistic regression; NN, neural network; PROMs, patient-reported outcome measures; QLO-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest; SVMs, support vector machines.

1. Original (no preprocessing method, 80% training set, 20% testing set)
2. Balanced (random sampling with replacement [oversampling] before train/test split with 8:2 ratio)
3. Partially balanced (train/test split with 8:2 ratio, stratification ensuring the same proportion of classes in both sets,²⁶ oversampling performed on the training set, leaving testing set imbalanced)

ML Model Development

Six ML models, namely, logistic regression (LR), decision tree (DT), adaptive boosting (AB), random forest (RF), support vector machines (SVMs), and neural network (NN), were selected on the basis of their inclusions in previous research.^{18,37} Hyperparameter tuning was performed on training sets through grid search with five-fold cross-validation.³⁷ The models were applied using Python *sklearn* library.

Model Evaluation

Accuracy, precision, recall (also known as sensitivity), F1 score, and AUC were used to evaluate model performance. AUC, a commonly used metric in ML studies, was considered a main metric for model evaluation to enable between-studies comparisons. Model calibrations were evaluated with calibration plots of RF in balanced data sets and LR in remaining data sets because of the best overall performance of these models in these scenarios. Feature importance analyses were also performed on these models. LR features were analyzed through the absolute values of regression coefficients. This method is only meaningful for standardized data with no multicollinearity,³⁸ which was accounted for by standardization of features and removing correlated variables. RF features were explored through “feature importances” python command in *sklearn.RandomForestClassifier*. Analysis of variance with Tukey’s honest significant difference tests were performed to compare model performances (Appendix Table A3).

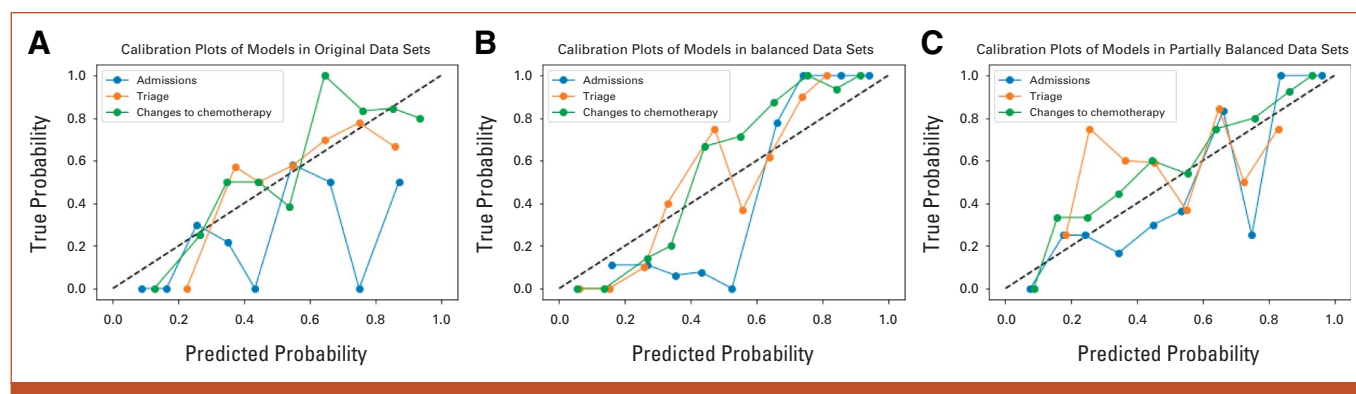


FIG 3. Calibration plots of LR models in original dataset (A), calibration of RF models in balanced datasets (B), and calibration of LR models in partially balanced datasets (C). LR, logistic regression; RF, random forest.

Clinical Feedback

Involving health professionals in early stages of exploratory research could support the clinical adoption of ML models.³⁹ Therefore, the methodology design was finalized after feedback from the Patient Centred Outcomes Research Group in the University of Leeds, Faculty of Medicine and Health, Leeds Institute of Medical Research. This group includes oncologists, nurses, and psychologists.

RESULTS

Performance metrics and hyperparameters for all models applied to all feature sets for all preprocessing methods are presented in Appendix Table A4.

Hospital Admissions

Overall Predictive Value of PROMs

For all models in original and balanced data sets, clinical variables had worse AUC than feature sets including PROMs. In the partially balanced data set, F1 score was higher for clinical variables in SVM (0.188) and NN (0.493) than for other feature sets. Nevertheless, these values were not the highest overall. For SVM, recall was also the highest value for the clinical variables (0.176). No AUC was the highest for clinical variables.

Predictive Value of Individual PROM Questionnaires

In the original data set, clinical + QLQ-C30 variables achieved the best AUC in all models except NN (AUC was highest for PROMs only). In the balanced data set, clinical + QLQ-C30 variables obtained the highest AUC in all models, apart from LR and SVM (AUCs were highest for all variables). In the partially balanced data set, the highest AUC was obtained by clinical + QLQ-C30 variables in LR, DT, RF (the same value as PROMs only), and NN. In AB, the highest AUC was achieved by all variables, and in SVM by

clinical + FACT-G variables. Therefore, QLQ-C30 variables aided ML performance the most.

Model Performance

In the original data set, LR performed best in all feature sets, except for clinical + EQ-5D, where DT was superior (Fig 2A). The highest AUC (0.659) was obtained by LR with clinical + QLQ-C30 variables. In the balanced data set, RF was the best performing algorithm (highest AUC = 0.905) for all feature sets, apart from all variables and clinical + EQ-5D variables, where SVM performed slightly better. In the partially balanced data set, the best AUC (0.616) was achieved by LR using clinical + QLQ-C30 variables and AB using all variables. Balancing the entire data set improved model performance on the basis of all evaluation metrics (Figs 2B and 2E). Using partially balanced data resulted in similar AUCs and precision to original data, but improved F1 score and recall. Model calibration for predicting admissions is poor in original data and improves slightly in balanced and partially balanced scenarios (Fig 3). LR prioritized clinical variables, while RF focused on PROMs and age at study entry (Table 1).

Triage Events

Overall Predictive Value of PROMs

No AUC was highest for clinical variables in any of the models and data sets, suggesting that PROMs improved model performance. The only highest values obtained by only clinical variables were F1 score and recall in original data (NN) precision in balanced data (AB), but these were not the highest values considering all models.

Predictive Value of Individual PROM Questionnaires

In the original data set, feature sets achieving the highest AUCs were only PROM variables for LR, DT, and RF; all variables for SVM and NN; and clinical + QLQ-C30 variables for AB. In the balanced data set, all variables obtained the

TABLE 1. Feature Importance for LR and RF Models Predicting All Three Target Variables

Target Variable	Type of Variable	Input Variable	Original Data				Balanced Data				Balanced Train/Real Test			
			LR		RF		LR		RF		LR		RF	
			Coefficient	Rank	Value	Rank	Coefficient	Rank	Value	Rank	Coefficient	Rank	Value	Rank
Admissions	Clinical/ demographic	AgeStudyEntry	0.005	19	0.052	5	−0.041	13	0.093	3	0.019	12	0.108	1
		StudyArm	0.060	11	0.006	23	0.100	12	0.013	22	0.085	9	0.021	20
		DiseaseSite	−0.348	4	0.019	17	−0.412	2	0.033	13	−0.404	3	0.032	15
		PreviousChemo	−0.402	2	0.003	24	−0.290	4	0.012	24	−0.241	4	0.012	23
		PrimaryorMet	0.866	1	0.037	13	1.047	1	0.044	9	0.732	1	0.033	12
		Comorbidities	0.366	3	0.034	14	0.335	3	0.039	10	0.484	2	0.055	6
		DaysonStudy	0.007	17	0.037	12	0.004	19	0.019	20	0.006	18	0.013	22
	QLQ-C30	C30_Appetite_0	0.008	16	0.106	3	0.011	15	0.050	5	0.008	16	0.036	10
		C30_Dyspnoea_0	0.013	14	0.128	2	0.013	14	0.048	7	0.017	14	0.042	8
		C30_NauseaVom_0	0.013	15	0.052	6	−0.003	24	0.017	21	0.006	19	0.017	21
		C30_Const_0	−0.003	20	0.009	21	−0.005	17	0.026	17	−0.002	23	0.033	14
		C30_Diarr_0	−0.006	18	0.013	20	−0.003	20	0.023	19	−0.004	21	0.024	19
		C30_Financ_0	−0.001	22	0.040	10	−0.007	16	0.030	15	−0.007	17	0.032	16
		C30_Cognitive_0	−0.001	23	0.030	15	−0.003	22	0.045	8	0.003	22	0.043	7
		C30_Sleep_0	−0.003	21	0.014	19	−0.003	21	0.035	12	−0.005	20	0.035	11
		C30_Social_0	0.000	24	0.039	11	0.005	18	0.048	6	−0.013	15	0.061	5
	EQ-5D	QoLEQ5DMob	−0.061	10	0.017	18	0.225	8	0.029	16	−0.002	24	0.030	17
		QoLEQ5DSelCar	0.197	6	0.055	4	0.112	11	0.013	23	−0.217	5	0.009	24
		QoLEQ5DUsuAct	0.165	8	0.043	9	−0.147	9	0.024	18	0.124	7	0.033	13
		QoLEQ5DPain	0.030	13	0.045	8	0.128	23	0.031	14	−0.132	6	0.029	18
		QoLEQ5DAnxDep	−0.039	12	0.009	22	0.239	7	0.037	11	−0.117	8	0.038	9
	FACT-G	PhysicalWB_Baseline	−0.188	7	0.132	1	−0.260	5	0.104	1	−0.017	13	0.094	2
		FunctionalWB_Baseline	0.163	9	0.051	7	0.126	10	0.102	2	−0.079	10	0.090	3
		SocialWB_Baseline	0.248	5	0.028	16	0.259	6	0.086	4	0.053	11	0.080	4
Triage	Clinical/ demographic	AgeStudyEntry	−0.127	10	0.097	2	−0.186	8	0.119	1	−0.265	3	0.121	1
		StudyArm	−0.162	9	0.006	23	−0.202	7	0.014	22	0.243	4	0.026	17
		DiseaseSite	−0.166	8	0.013	22	−0.255	5	0.027	16	−0.196	7	0.030	15
		PreviousChemo	−0.721	1	0.019	19	−0.430	3	0.013	24	−0.643	1	0.022	20
		PrimaryorMet	0.487	2	0.004	24	0.644	1	0.018	20	0.624	2	0.017	23
		Comorbidities	0.338	5	0.044	10	0.397	4	0.036	11	0.108	9	0.036	10
		DaysonStudy	0.012	15	0.034	12	0.001	24	0.014	23	0.010	19	0.020	21
	QLQ-C30	C30_Appetite_0	0.006	18	0.077	3	0.002	23	0.032	12	0.000	24	0.031	14
		C30_Dyspnoea_0	0.003	23	0.028	16	0.004	21	0.026	17	0.015	16	0.019	22
		C30_NauseaVom_0	−0.005	20	0.031	13	0.009	16	0.019	19	0.002	22	0.022	19
		C30_Const_0	−0.006	17	0.016	20	−0.010	15	0.032	13	−0.017	15	0.034	13
		C30_Diarr_0	0.004	21	0.053	8	0.008	17	0.029	14	0.026	12	0.036	9
		C30_Financ_0	−0.004	22	0.019	18	−0.006	19	0.026	18	−0.010	20	0.029	16
		C30_Cognitive_0	−0.005	19	0.062	6	0.002	22	0.043	7	−0.009	21	0.044	7
		C30_Sleep_0	0.006	16	0.050	9	0.005	20	0.049	5	0.012	18	0.051	5
		C30_Social_0	−0.002	24	0.029	15	−0.008	18	0.049	6	−0.001	23	0.044	6
	EQ-5D	QoLEQ5DMob	0.024	13	0.013	21	−0.183	10	0.028	15	−0.115	8	0.026	18
		QoLEQ5DSelCar	−0.444	3	0.042	11	−0.207	6	0.014	21	−0.222	6	0.008	24
		QoLEQ5DUsuAct	0.283	6	0.056	7	0.088	14	0.037	9	0.021	13	0.034	12
		QoLEQ5DPain	0.053	12	0.030	14	0.440	2	0.040	8	0.019	14	0.035	11
		QoLEQ5DAnxDep	0.069	11	0.019	17	0.100	13	0.037	10	0.013	17	0.037	8
	FACT-G	PhysicalWB_Baseline	−0.258	7	0.122	1	−0.131	12	0.096	4	−0.234	5	0.091	3
		FunctionalWB_Baseline	0.364	4	0.074	4	0.186	9	0.103	2	0.074	10	0.096	2
		SocialWB_Baseline	0.013	14	0.063	5	0.171	11	0.099	3	0.036	11	0.090	4
(continued on following page)														

TABLE 1. Feature Importance for LR and RF Models Predicting All Three Target Variables (continued)

Target Variable	Type of Variable	Input Variable	Original Data				Balanced Data				Balanced Train/Real Test			
			LR		RF		LR		RF		LR		RF	
			Coefficient	Rank	Value	Rank	Coefficient	Rank	Value	Rank	Coefficient	Rank	Value	Rank
Changes to chemotherapy	Clinical/ demographic	AgeStudyEntry	−0.295	6	0.113	2	−0.138	6	0.112	2	−0.179	9	0.127	1
		StudyArm	−0.167	9	0.008	24	−0.137	7	0.016	22	−0.084	13	0.021	19
		DiseaseSite	1.007	1	0.128	1	−0.197	4	0.098	3	0.755	1	0.071	5
		PreviousChemo	0.034	12	0.012	23	−0.159	5	0.016	21	−0.226	6	0.015	22
		PrimaryorMet	−0.723	2	0.022	17	0.294	3	0.018	19	−0.299	3	0.018	20
		Comorbidities	0.572	3	0.064	6	0.316	2	0.031	12	0.396	2	0.038	10
		DaysonStudy	−0.002	21	0.024	16	0.000	24	0.014	23	0.003	23	0.009	24
	QLQ-C30	C30_Appetite_0	0.010	17	0.035	8	0.003	22	0.039	8	0.011	17	0.029	14
		C30_Dyspnoea_0	0.000	24	0.020	19	0.004	21	0.013	24	0.005	22	0.016	21
		C30_NauseaVom_0	0.009	18	0.019	21	0.008	16	0.021	18	0.037	15	0.021	19
		C30_Const_0	0.004	20	0.027	14	−0.010	15	0.026	14	−0.008	20	0.029	15
		C30_Diarr_0	−0.010	16	0.022	18	0.008	17	0.023	16	−0.011	18	0.028	16
		C30_Financ_0	−0.010	15	0.029	12	−0.006	19	0.032	11	−0.009	19	0.034	11
		C30_Cognitive_0	0.000	23	0.024	15	0.001	23	0.038	9	0.014	16	0.044	7
		C30_Sleep_0	0.001	22	0.044	7	0.005	20	0.040	7	0.006	21	0.041	9
		C30_Social_0	0.006	19	0.034	9	−0.007	18	0.046	6	−0.001	24	0.047	6
	EQ-5D	QoLEQ5DMob	0.241	7	0.033	10	−0.135	9	0.027	13	0.137	10	0.027	17
		QoLEQ5DSelCar	0.446	4	0.019	20	−0.129	10	0.017	20	0.282	4	0.013	23
		QoLEQ5DUsuAct	0.052	11	0.032	11	0.043	14	0.034	10	0.218	7	0.041	8
		QoLEQ5DPain	−0.211	8	0.029	13	0.334	1	0.024	15	−0.128	11	0.031	13
		QoLEQ5DAnxDep	−0.136	10	0.016	22	0.063	13	0.023	17	−0.056	14	0.031	12
	FACT-G	PhysicalWB_Baseline	−0.011	14	0.066	5	−0.125	11	0.064	5	0.274	5	0.079	4
		FunctionalWB_Baseline	−0.026	13	0.112	3	0.108	12	0.143	1	−0.091	12	0.109	2
		SocialWB_Baseline	−0.358	5	0.071	4	0.136	8	0.085	4	−0.215	8	0.082	3

Abbreviations: EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; LR, logistic regression; QLQ-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest.

highest AUC for DT, AB, RF, and SVM. Clinical + QLQ-C30 variables had the highest performance for LG, and clinical + EQ-5D variables for NN. In the partially balanced data set, clinical + EQ-5D variables had the best performance the most frequently (for AB, RF, SVM, and NN), and clinical + FACT-G variables were selected twice (for LR and DT).

Model Performance

Overall, models predicting triage performed significantly worse than models predicting changes to chemotherapy ($P < .001$) and admissions ($P < .01$). LR outperformed other models, achieving the highest AUC values across all feature sets, apart from PROMs only variables in original data, where DT achieved the best AUC. In the balanced data set, SVM and RF performed the best (highest AUC = 0.764 for SVM for all variables). In the partially balanced data set, the best AUC (0.624) was obtained by LR in the clinical + FACT-G feature set. There was no outstanding model in the original data set, but in balanced and partially balanced data sets, SVM, RF, and NN provided the best F1 scores (Figs 4E and 4F). A slight increase in the AUCs for the balanced data set is noticeable in Figure 4B. AUCs and F1 scores in the partially balanced data

set were similar to the original data set. For some models in balanced data, the F1 scores were lower than in other data sets. Model calibration remained poor across different data types (Fig 3). LR's main features comprised clinical data with individual PROM variables, while RF primarily considered FACT-G variables (Table 1).

Changes to Chemotherapy

Overall Predictive Value of PROMs

In the original data set, AUC was the highest for clinical variables only in DT (0.623) and RF (0.623). However, these values were not the highest overall. In the balanced data set, recall was the only measure highest for clinical variables in AB (0.754) and NN (1). In the partially balanced data set, clinical variables obtained the highest precision (0.872) and AUC (0.682) in LR and SVM, respectively. Overall, highest AUC had models including PROMs.

Predictive Value of Individual PROM Questionnaires

In the original data set, clinical + QLQ-C30 variables had the highest performance in LR and SVM; clinical + FACT-G

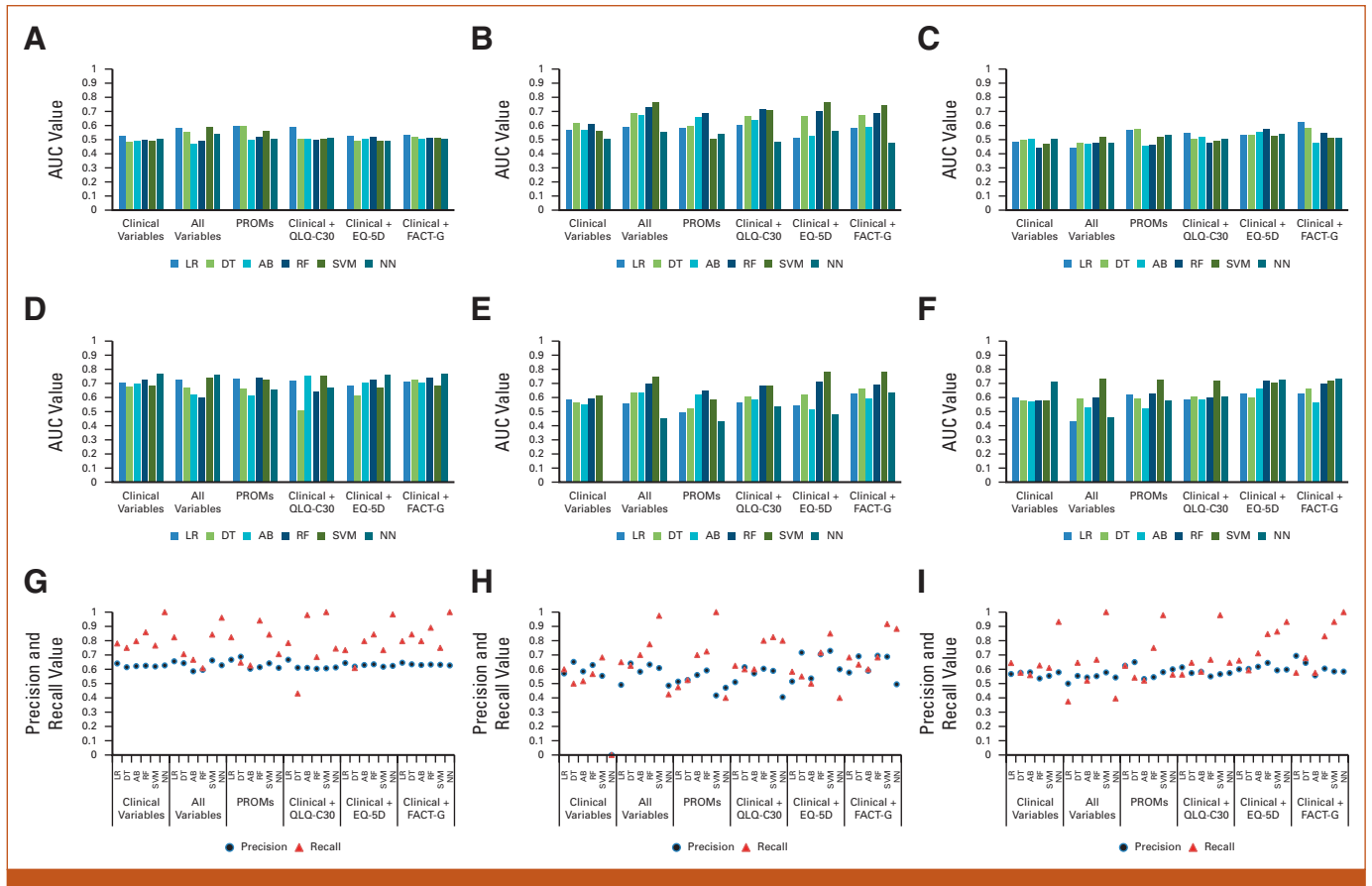


FIG 4. Predicting triage. AUC values for (A) original, (B) balanced, and (C) partially balanced data sets obtained from all models in all feature sets; F1 scores for (D) original, (E) balanced, and (F) partially balanced data sets obtained from all models in all feature sets; precision and recall distributions for (G) original, (H) balanced, and (I) partially balanced data sets obtained from all models in all feature sets. AB, adaptive boosting; DT, decision tree; EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; LR, logistic regression; NN, neural network; PROMs, patient-reported outcome measures; QLQ-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest; SVMs, support vector machines.

variables for AB and NN; and clinical variables for DT and RF. In the balanced data set, clinical + QLQ-C30 variables obtained the highest AUC in DT and AB, clinical + EQ-5D variables for NN, and all variables for LR and RF. In SVM, all variables and clinical + QLQ-C30 variables achieved the same AUC (0.931). In the partially balanced data set, all variables obtained the highest AUC for DT, AB, and RF; only clinical variables for LR and SVM; and clinical + EQ-5D variables for NN.

Model Performance

Overall, models predicting changes to chemotherapy performed significantly better than models predicting triage ($P < .01$) and admissions ($P < .001$). No model in original and partially balanced data sets outperformed others. In the balanced data set, the best algorithms were RF and SVM. SVM with all and clinical + QLQ-C30 variables had the best overall performance (AUC = 0.931). The AUCs of the models increased when data were balanced, but there was no difference in F1 scores. There was no noticeable difference between original and partially balanced data sets (Figs 5D and 5F). Model calibration was very good in the partially

balanced data set and slightly worse in other data types (Fig 3). LR prioritized clinical variables with individual FACT-G and EQ-5D features. RF mainly considered FACT-G and some clinical variables (Table 1).

DISCUSSION

We successfully applied a range of ML models to a complex oncology data set with clinical, PROM, and health outcome data. PROMs improved the overall performance of ML models for all target variables. Sometimes the best performing models included only PROM variables. Although there is evidence suggesting that using PROMs without objectively measured data in ML models can lead to accurate predictions,¹⁵ this study encourages using both clinical and PROM data. The QLQ-C30 questionnaire added the most predictive value overall. This might be explained by QLQ-C30 being the only questionnaire with variables consistently significantly different between classes. These results are promising, as the wide availability of QLQ-C30⁴⁰ may aid its utilization in ML models for clinical practice.

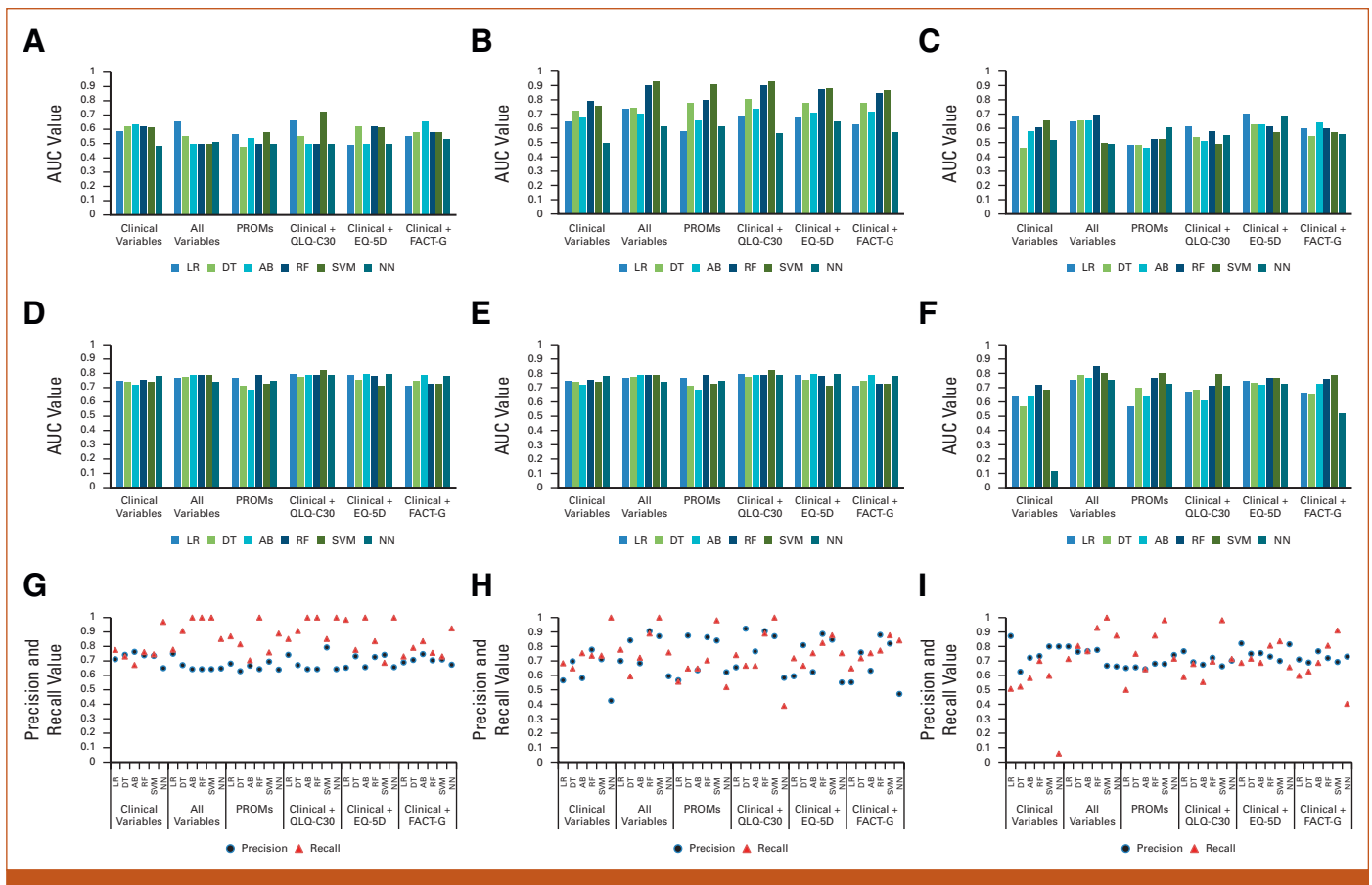


FIG 5. Predicting changes to chemotherapy. AUC values for (A) original, (B) balanced, and (C) partially balanced data sets obtained from all models in all feature sets predicting; F1 scores for (D) original, (E) balanced, and (F) partially balanced data sets obtained from all models in all feature sets predicting; precision and recall distributions for (G) original, (H) balanced, and (I) partially balanced data sets obtained from all models in all feature sets. AB, adaptive boosting; DT, decision tree; EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; LR, logistic regression; NN, neural network; PROMs, patient-reported outcome measures; QOL-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest; SVMs, support vector machines.

LR being the simplest model and outperforming other methods in imbalanced data was also observed in previous studies.^{26,35} Good performance of RF and SVM when predicting admissions and changes to chemotherapy in balanced data is compatible with ensemble methods of previously reported outcome predictions.^{17,18,41} Changes to chemotherapy predictions had the best overall performance, which is further confirmed by great calibration of models predicting this target in the partially balanced data set. This might be explained by more frequent and stronger significance of feature differences between classes. Poor performance of triage predictions could be due to more subjective nature of this target, compared with clinical decision to admit a patient or make treatment changes. Balancing data sets improved overall model performance. Using the balanced data set might decrease generalizability of models, as oversampling often causes overfitting.⁴² Therefore, evaluating models on the balanced testing set prevents the models from applications in clinical practice, as the real-world data are never perfectly balanced. Nevertheless, training models on imbalanced data can lead to incorrect predictions, biased

toward one of the classes,⁴³ which was apparent through low recall in admission predictions, being the most imbalanced target. Using the partially balanced data set mitigates such bias and the lack of generalizability. This method ensures robustness of models through the balanced training set and obtains a more accurate perspective for real clinical data through the original testing set.⁴⁴

In all target variables, LR models focused more on clinical features than PROMs. RF models usually favored FACT-G variables with some relevant clinical or QOL-C30 features (mainly sleep, cognitive, and social scales). Although these patterns were similar for all target variables, changes to chemotherapy predictions showed the smallest discrepancy between the feature ranks. It might be explained by the best predictive performance of this outcome. LR was often the best performing model in original data, which could explain its poor performance of predicting triage and admissions, as the clinical features for these targets did not have significant differences between classes (Appendix Table A3), yet the model was considering these variables the most important

(Table 1). For changes to chemotherapy, there were many significantly different clinical features, explaining good performance of LR. RF usually favored PROMs, which explained this model struggling to predict outcomes from only clinical variables.

Inclusion of three different PROMs allowed understanding of their individual predictive value. This study addressed the inconsistency in preprocessing methods for class imbalance in existing studies^{18,33,35–37} and highlighted differences in results generated from these three techniques. The variety of performance metrics reported allowed between-studies comparison¹⁵ and in-depth understanding of models. Furthermore, consulting clinicians during study design ensured clinical relevance of research questions, which can support adoption of ML methods.³⁹

The limitations of this study include clinical trial data collection, which might not be representative of the population.⁴⁵ No information about patients' ethnicity was provided, which limited understanding of potential bias in data⁴⁶ and prevented subgroup analyses.³³ Small sample size is associated with higher accuracy in classification,⁴⁷ so using more data would prevent potential bias. Furthermore, this work used only PROMs collected at the beginning of chemotherapy (baseline), so potential over-time

dependencies of patient reports were missed. Half of the participants used clinical trial intervention, which might have affected the outcome, but this risk was mitigated through performance comparison in control and intervention groups, identifying no significant differences.

In conclusion, this study supported the evidence that PROMs, such as health-related quality of life, functioning, and symptom reporting, can improve the performance of ML models predicting patient outcomes. The predictive value of widely available PROMs, such as the QLQ-C30 questionnaire, supports the motivation for collecting and using these measures in ML research. The results inform further exploration of PROMs' effect as predictors, and potential application of ML models in clinical practice, if rigorous justification and reporting of methodology is performed. On the basis of large discrepancies across results from different preprocessing methods, this research alerts scientific community to justify choices on the methods for balancing data. It is recommended to balance the training set only and to test models on original data to prevent bias. In future work, we plan to involve patients and clinicians to assess their attitudes to ML-based prediction and to explore the broader implications of the findings. We also plan to use PROM data collected longitudinally throughout chemotherapy treatment, as over-time changes in reporting might provide more meaningful conclusions.

AFFILIATIONS

¹UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care, University of Leeds, Leeds, United Kingdom

²School of Computing, University of Leeds, Leeds, United Kingdom

³Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, United Kingdom

⁴Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom

⁵Leeds Institute of Health Sciences, University of Leeds, Leeds, United Kingdom

CORRESPONDING AUTHOR

Zuzanna Wójcik, MSc; e-mail: sczw@leeds.ac.uk.

DISCLAIMER

The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research (NIHR) or the Department of Health and Social Care.

PRIOR PRESENTATION

Presented in part at the 7th UK Patient Reported Outcome Measures (PROMs) Research Conference, "PROMs Across the Lifespan," Sheffield, UK, June 22, 2023, and at the Joint Centre for Doctoral Training Conference on Artificial Intelligence for Healthcare (CAI4H), York, UK, May 4-5, 2023.

SUPPORT

Supported in part by UK Research and Innovation (UKRI; CDT grant number EP/S024336/1). The data set used in the analysis was collected in a study funded by the National Institute for Health and Care

Research under its Program Grants for Applied Research (Ref: NIHR RP-PG-0611-20008).

AUTHOR CONTRIBUTIONS

Conception and design: Zuzanna Wójcik, Vania Dimitrova, Galina Velikova, Kate Absolom

Collection and assembly of data: Lorraine Warrington, Galina Velikova, Kate Absolom

Data analysis and interpretation: All authors

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted.

I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](http://OpenPayments)).

Galina Velikova

Honoraria: Eisai, Pfizer, Novartis

Consulting or Advisory Role: Roche UK, Eisai, Novartis, Sanofi, Pfizer, AstraZeneca, Seagen

Speakers' Bureau: Novartis

Research Funding: Pfizer (Inst), IQVIA (Inst)

Travel, Accommodations, Expenses: Roche UK, Novartis, Eisai

Other Relationship: University of Leeds

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

The authors thank the Patient Centred Outcomes Research (PCOR) Group in the University of Leeds, Faculty of Medicine and Health, Leeds Institute of Medical Research for consultations of the applicability of

the study methodology. The authors also thank the patients and clinicians participating in the eRAPID clinical trial for providing the data. The authors also thank the patient representative of the UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care Richard Stephens for discussing the study design and its relevance to patients, professor David Hogg and Dr Amy Downing for their feedback on an early draft, and Dr Sam Relton for his advice on the statistical analysis.

REFERENCES

- Du XL, Osborne C, Goodwin JS: Population-based assessment of hospitalizations for toxicity from chemotherapy in older women with breast cancer. *J Clin Oncol* 20:4636-4642, 2002
- Amjad MT, Chidhara A, Kasi A: *Cancer Chemotherapy*. Treasure Island, FL, StatPearls Publishing, 2022
- Livingston PM, Craike M, Considine J: Unplanned presentations to emergency departments due to chemotherapy induced complications: Opportunities for improving service delivery. *Aust Emerg Nurs J* 14:62-68, 2011
- Dufton PH, Gerditz MF, Jarden R, et al: Methodological approaches to measuring the incidence of unplanned emergency department presentations by cancer patients receiving systemic anti-cancer therapy: A systematic review. *BMC Med Res Methodol* 22:75, 2022
- Hong JC, Niedzwiecki D, Palta M, et al: Predicting emergency visits and hospital admissions during radiation and chemoradiation: An internally validated pretreatment machine learning algorithm. *JCO Clin Cancer Inform* 10.1200/CCI.18.00037
- Feliu J, Espinosa E, Basterretxea L, et al: Prediction of chemotoxicity, unplanned hospitalizations and early death in older patients with colorectal cancer treated with chemotherapy. *Cancers* 14:127, 2021
- Yang CC: Explainable artificial intelligence for predictive modeling in healthcare. *J Healthc Inform Res* 6:228-239, 2022
- Zeng S, Li L, Hu Y, et al: Machine learning approaches for the prediction of postoperative complication risk in liver resection patients. *BMC Med Inform Decis Mak* 21:371, 2021
- Harari Y, O'Brien MK, Lieber RL, et al: Inpatient stroke rehabilitation: Prediction of clinical outcomes using a machine-learning approach. *J Neuroeng Rehabil* 17:71, 2020
- Abbasi B, Goldenholz DM: Machine learning applications in epilepsy. *Epilepsia* 60:2037-2047, 2019
- Thorsen-Meyer HC, Nielsen AB, Nielsen AP, et al: Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health* 2:e179-e191, 2020
- Hong WS, Haimovich AD, Taylor RA: Predicting hospital admission at emergency department triage using machine learning. *PLoS One* 13:e0201016, 2018
- Kingsley C, Patel S: Patient-reported outcome measures and patient-reported experience measures. *BJA Educ* 17:137-144, 2017
- Tschuggnall M, Grote V, Pirchl M, et al: Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Inform Med Unlocked* 24:100598, 2021
- Verma D, Bach K, Mork PJ: Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review. *Informatics* 8:56, 2021
- Eton DT, Fairclough DL, Cella D, et al: Early change in patient-reported health during lung cancer chemotherapy predicts clinical outcomes beyond those predicted by baseline report: Results from Eastern Cooperative Oncology Group Study 5592. *J Clin Oncol* 21:1536-1543, 2003
- Sidey-Gibbons C, Pfof A, Asaad M, et al: Development of machine learning algorithms for the prediction of financial toxicity in localized breast cancer following surgical treatment. *JCO Clin Cancer Inform* 10.1200/CCI.20.00088
- Sim J-A, Kim YA, Kim JH, et al: The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: Applications of machine learning. *Sci Rep* 10:10693, 2020
- Kumar V, Roche C, Overman S, et al: What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? *Clin Orthop Relat Res* 478: 2351-2363, 2020
- Sun R, Tomkins-Lane C, Muaremi A, et al: Physical activity thresholds for predicting longitudinal gait decline in adults with knee osteoarthritis. *Osteoarthritis Cartilage* 29:965-972, 2021
- Tasci E, Zhuge Y, Camphausen K, et al: Bias and class imbalance in oncologic data—Towards inclusive and transferrable AI in large scale oncology data sets. *Cancers* 14:2897, 2022
- Absolom K, Warrington L, Hudson E, et al: Phase III randomized controlled trial of eRAPID: eHealth intervention during chemotherapy. *J Clin Oncol* 39:734-747, 2021
- Aaronson NK, Ahmedzai S, Bergman B, et al: The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 85:365-376, 1993
- Dolan P: Modeling valuations for EuroQol health states. *Med Care* 35:1095-1108, 1997
- Cella DF, Tulsky DS, Gray G, et al: The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. *J Clin Oncol* 11:570-579, 1993
- Munn JS, Lanting BA, MacDonald SJ, et al: Logistic regression and machine learning models cannot discriminate between satisfied and dissatisfied total knee arthroplasty patients. *J Arthroplasty* 37:267-273, 2022
- Harrison CJ, Geoghegan L, Sidey-Gibbons CJ, et al: Developing machine learning algorithms to support patient-centered, value-based carpal tunnel decompression surgery. *Plast Reconstr Surg Glob Open* 10:e4279, 2022
- Wiemken TL, Kelley RR: Machine learning in epidemiology and health outcomes research. *Annu Rev Public Health* 41:21-36, 2020
- Wshah S, Skalka C, Price M: Predicting posttraumatic stress disorder risk: A machine learning approach. *JMIR Ment Health* 6:e13946, 2019
- Goh L, Song Q, Kasabov N: A novel feature selection method to improve classification of gene expression data. *Second Asia-Pacific Bioinformatics Conference (APBC 2004)*, Dunedin, New Zealand, January 18-22, 2004
- Bennett DA: How can I deal with missing data in my study? *Aust N Z J Public Health* 25:464-469, 2001
- Haneuse S, Arterburn D, Daniels MJ: Assessing missing data assumptions in EHR-based studies: A complex and underappreciated task. *JAMA Netw Open* 4:e210184, 2021
- Pfof A, Mehrara BJ, Nelson JA, et al: Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPIRED-001). *Breast* 60:111-122, 2021
- Kober KM, Roy R, Dhruva A, et al: Prediction of evening fatigue severity in outpatients receiving chemotherapy: Less may be more. *Fatigue* 9:14-32, 2021
- Staartjes VE, de Wispelaere MP, Vandertop WP, et al: Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: Feasibility of center-specific modeling. *Spine J* 19:853-861, 2019
- Zhang S, Chen JY, Pang HN, et al: Development and internal validation of machine learning algorithms to predict patient satisfaction after total hip arthroplasty. *Arthroplasty* 3:33, 2021
- Sidey-Gibbons CJ, Sun C, Schneider A, et al: Predicting 180-day mortality for women with ovarian cancer using machine learning and patient-reported outcome data. *Sci Rep* 12:21269, 2022
- Cava WL, Bauer C, Moore JH, et al: Interpretation of machine learning predictions for patient outcomes in electronic health records. *AMIA Annu Symp Proc* 2019:572-581, 2019
- Davenport T, Kalakota R: The potential for artificial intelligence in healthcare. *Future Healthc J* 6:94-98, 2019
- Kim SH, Jo MW, Kim HJ, et al: Mapping EORTC QLQ-C30 onto EQ-5D for the assessment of cancer patients. *Health Qual Life Outcomes* 10:151, 2012
- Peterson DJ, Ostberg NP, Blayney DW, et al: Machine learning applied to electronic health records: Identification of chemotherapy patients at high risk for preventable emergency department visits and hospital admissions. *JCO Clin Cancer Inform* 10.1200/CCI.21.00116
- Azimi V, Zaydman MA: Optimizing equity: Working towards fair machine learning algorithms in laboratory medicine. *J Appl Lab Med* 8:113-128, 2023
- Shafique R, Rustam F, Choi GS, et al: Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning. *Cancers* 15:681, 2023
- Siccoli A, de Wispelaere MP, Schröder ML, et al: Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg Focus* 46:E5, 2019
- Blonde L, Khunti K, Harris SB, et al: Interpretation and impact of real-world clinical data for the practicing clinician. *Adv Ther* 35:1763-1774, 2018
- Ghassemi M, Nsoesie EO: In medicine, how do we machine learn anything real? *Patterns* 3:100392, 2022
- Vabalas A, Gowen E, Poliakov E, et al: Machine learning algorithm validation with a limited sample size. *PLoS One* 14:e0224365, 2019

APPENDIX

TABLE A1. Variables in Each Feature Sets

Clinical Variable	All Variables	PROMs	Clinical + QLQ-C30	Clinical + EQ-5D	Clinical + FACT-G
DiseaseSite	DiseaseSite	PhysicalWB Baseline	DiseaseSite	DiseaseSite	DiseaseSite
Sex	Sex	SocialWB Baseline	Sex	Sex	Sex
PreviousChemo	PreviousChemo	EmotionalWB Baseline	PreviousChemo	PreviousChemo	PreviousChemo
AgeStudyEntry	AgeStudyEntry	FunctionalWB Baseline	AgeStudyEntry	AgeStudyEntry	AgeStudyEntry
PrimaryorMet	PrimaryorMet	QLQ-C30 Appetite 0	PrimaryorMet	PrimaryorMet	PrimaryorMet
Comorbidities	Comorbidities	QLQ-C30 Dyspnoea 0	Comorbidities	Comorbidities	Comorbidities
DaysonStudy	DaysonStudy	QLQ-C30 Pain 0	DaysonStudy	DaysonStudy	DaysonStudy
StudyArm	StudyArm	QLQ-C30 Fatigue 0	StudyArm	StudyArm	StudyArm
	PhysicalWB Baseline	QLQ-C30 NauseaVom 0	QLQ-C30 Appetite 0	QoLEQ5DMob	PhysicalWB Baseline
	SocialWB Baseline	QLQ-C30 Const 0	QLQ-C30 Dyspnoea 0	QoLEQ5DSelCar	SocialWB Baseline
	EmotionalWB Baseline	QLQ-C30 Diarr 0	QLQ-C30 Pain0	QoLEQ5DUusuAct	FunctionalWB Baseline
	EmotionalWBBaseline	QLQ-C30 Financ0	QLQ-C30 Fatigue 0	QoLEQ5DPain	
	FunctionalWB Baseline	QLQ-C30 GlobalHealth 0	QLQ-C30 NauseaVom 0QLQ-C30	QoLEQ5DAnxDep	
	QLQ-C30 Appetite 0	QLQ-C30 Cognitive 0	Const 0		
	QLQ-C30 Dyspnoea 0	QLQ-C30 Sleep 0	QLQ-C30 Diarr 0		
	QLQ-C30 Pain 0	QLQ-C30 Emotional 0	QLQ-C30 Financ 0		
	QLQ-C30 Fatigue 0	QLQ-C30 Physical 0	QLQ-C30 GlobalHealth 0		
	QLQ-C30 NauseaVom 0	QLQ-C30 Role 0	QLQ-C30 Cognitive 0		
	QLQ-C30 Const 0	QLQ-C30 Social 0	QLQ-C30 Sleep 0		
	QLQ-C30 Diarr 0	QoLEQ5DMob	QLQ-C30 Emotional 0		
	QLQ-C30 Financ 0	QoLEQ5DSelCar	QLQ-C30 Physical 0		
	QLQ-C30 GlobalHealth 0	QoLEQ5DUusuAct	QLQ-C30 Role 0		
	QLQ-C30 Cognitive 0	QoLEQ5DPain	QLQ-C30 Social 0		
	QLQ-C30 Sleep 0	QoLEQ5DAnxDep			
	QLQ-C30 Emotional 0				
	QLQ-C30 Physical 0				
	QLQ-C30 Role 0				
	QLQ-C30 Social 0				
	QoLEQ5DMob				
	QoLEQ5DSelCar				
	QoLEQ5DUusuAct				
	QoLEQ5DPain				
	QoLEQ5DAnxDep				

NOTE. Variables in bold were selected in ML models.
Abbreviations: EQ-5D, EuroQol Five-dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; PROMs, patient-reported outcome measures; QLQ-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire.

TABLE A2. Differences in Features Between Classes in All Target Variables and the *P* Value Generated by Using Mann-Whitney *U* Test for All Features but AgeStudyEntry (*t*-test used), as It Was the Only Normally Distributed Variable

Input Variable	Not Admitted (n = 338)		Admitted (n = 170)		Significance Test	
	Median	Mean/Value	Median	Mean/Value	<i>P</i>	Significance Level
StudyArm	1	1 = 170, 2 = 168	1	1 = 86, 2 = 84	.475	
DiseaseSite	2	1 = 157, 2 = 58, 3 = 123	2	1 = 76, 2 = 48, 3 = 46	.220	
Sex	2	F = 263, M = 75	2	F = 143, M = 27	.047	*
PreviousChemo	0	No = 266, Yes = 72	0	No = 131, Yes = 39	.337	
AgeStudyEntry	56	55.429	57.5	56.994	.158	
PrimaryorMet	1	1 = 228, 2 = 110	1	1 = 89, 2 = 81	<.001	**
Comorbidities	0	0 = 192, 1 = 103, 2 = 36, 3 = 7	1	0 = 75, 1 = 54, 2 = 30, 3 = 11	<.001	**
DaysonStudy	126	119.891	126	119.888	.118	
PhysicalWB_Baseline	25	23.725	24	22.424	.003	***
SocialWB_Baseline	24	23.693	24	23.552	.126	
EmotionalWB_Baseline	18	16.769	17	16.429	.131	
FunctionalWB_Baseline	19.917	19.016	19	18.263	.123	
C30_Appetite_0	0	16.606	0	27.381	.115	
C30_Dyspnoea_0	0	8.424	0	20.000	.003	***
C30_Pain_0	16.667	20.818	16.667	25.833	.124	
C30_Fatigue_0	22.222	26.835	33.333	36.984	.002	***
C30_NauseaVom_0	0	3.879	0	9.643	.028	*
C30_Const_0	0	12.774	0	15.952	.160	
C30_Diarr_0	0	9.179	0	10.000	.443	
C30_Financ_0	0	15.644	0	13.810	.246	
C30_GlobalHealth_0	75	73.853	75	67.679	.037	*
C30_Cognitive_0	83.333	83.514	83.333	80.595	.250	
C30_Sleep_0	33.333	37.681	33.333	37.619	.388	
C30_Emotional_0	75	73.345	83.333	75.060	.261	
C30_Physical_0	93.333	84.073	80	76.976	.012	*
C30_Role_0	83.333	73.980	83.333	70.513	.103	
C30_Social_0	83.333	74.527	83.333	72.189	.224	
QoLEQ5DMob	1	1.369	1	1.547	.032	*
QoLEQ5DSelCar	1	1.141	1	1.241	.056	
QoLEQ5DUsuAct	1	1.731	2	1.847	.107	
QoLEQ5DPain	2	1.810	2	1.947	.101	
QoLEQ5DAnxDep	2	1.807	2	1.894	.183	

Input Variable	Not Triaged (n = 214)		Triaged (n = 294)		Significance Test	
	Median	Mean/Value	Median	Mean/Value	<i>P</i>	Significance Level
StudyArm	2	1 = 104, 2 = 110	1	1 = 152, 2 = 142	.245	
DiseaseSite	2	1 = 85, 2 = 49, 3 = 80	1	1 = 148, 2 = 57, 3 = 89	.011	*
Sex	2	F = 162, M = 52	2	F = 244, M = 50	.022	*
PreviousChemo	0	No = 155, Yes = 59	0	No = 242, Yes = 52	.004	***
AgeStudyEntry	57	56.972	55	55.21	.097	
PrimaryorMet	1	1 = 132, 2 = 82	1	1 = 185, 2 = 109	.388	
Comorbidities	0	0 = 119, 1 = 62, 2 = 25, 3 = 8	0	0 = 148, 1 = 95, 2 = 41, 3 = 10	.136	
DaysonStudy	126	119.575	126	120.119	.489	
PhysicalWB_Baseline	25	23.711	24	22.980	.071	
SocialWB_Baseline	24	23.451	24	23.788	.305	
EmotionalWB_Baseline	18	16.872	18	16.497	.173	
FunctionalWB_Baseline	20	19.088	19	18.527	.154	

(continued on following page)

TABLE A2. Differences in Features Between Classes in All Target Variables and the *P* Value Generated by Using Mann-Whitney *U* Test for All Features but AgeStudyEntry (*t*-test used), as It Was the Only Normally Distributed Variable (continued)

Input Variable	Not Triaged (n = 214)		Triaged (n = 294)		Significance Test	
	Median	Mean/Value	Median	Mean/Value	<i>P</i>	Significance Level
C30_Appetite_0	0	17.326	0	22.361	.043	*
C30_Dyspnoea_0	0	10.358	0	13.750	.090	
C30_Pain_0	16.667	19.868	16.667	24.444	.057	
C30_Fatigue_0	22.222	26.962	33.333	32.662	.012	*
C30_NauseaVom_0	0	4.571	0	6.736	.012	*
C30_Const_0	0	15.048	0	12.971	.334	
C30_Diarr_0	0	7.910	0	10.600	.043	*
C30_Financ_0	0	15.443	0	14.722	.358	
C30_GlobalHealth_0	83.333	74.479	75	69.792	.122	
C30_Cognitive_0	83.333	85.217	83.333	80.556	.159	
C30_Sleep_0	33.333	35.028	33.333	39.609	.033	*
C30_Emotional_0	83.333	76.177	75	72.257	.227	
C30_Physical_0	86.667	82.420	86.667	81.153	.493	
C30_Role_0	83.333	75.708	83.333	70.719	.046	*
C30_Social_0	83.333	75.236	83.333	72.660	.234	
QoLEQ5DMob	1	1.458	1	1.409	.344	
QoLEQ5DSelCar	1	1.202	1	1.155	.313	
QoLEQ5DUsuAct	1	1.746	1	1.787	.239	
QoLEQ5DPain	2	1.829	2	1.876	.309	
QoLEQ5DAnxDep	2	1.762	2	1.890	.060	

Input Variable	No Chemotherapy Change (n = 175)		Chemotherapy Change (n = 333)		Significance Test	
	Median	Mean/Value	Median	Mean/Value	<i>P</i>	Significance Level
StudyArm	2	1 = 82, 2 = 93	1	1 = 174, 2 = 159	.124	
DiseaseSite	1	1 = 122, 2 = 21, 3 = 32	2	1 = 111, 2 = 85, 3 = 137	<.001	
Sex	2	F = 150, M = 25	2	F = 256, M = 77	.009	***
PreviousChemo	0	No = 145, Yes = 30	0	No = 252, Yes = 81	.032	*
AgeStudyEntry	54	54.126	58	56.913	.011	*
PrimaryorMet	1	1 = 128, 2 = 47	1	1 = 189, 2 = 144	<.001	**
Comorbidities	0	0 = 108, 1 = 52, 2 = 14, 3 = 1	1	0 = 159, 1 = 105, 2 = 52, 3 = 17	<.001	**
DaysonStudy	126	120.211	126	119.721	.085	
PhysicalWB_Baseline	25	23.822	24.5	23.015	.039	*
SocialWB_Baseline	25	24.325	24	23.295	<.001	**
EmotionalWB_Baseline	18	16.662	18	16.653	.346	
FunctionalWB_Baseline	20.5	19.572	19	18.346	.010	*
C30_Appetite_0	0	13.043	0	23.775	.041	*
C30_Dyspnoea_0	0	9.179	0	13.859	.291	
C30_Pain_0	16.667	21.981	16.667	22.760	.377	
C30_Fatigue_0	22.222	25.644	33.333	32.517	.089	
C30_NauseaVom_0	0	3.406	0	7.014	.205	
C30_Const_0	0	10.706	0	15.403	.337	
C30_Diarr_0	0	8.213	0	10.072	.379	
C30_Financ_0	0	16.667	0	14.217	.155	
C30_GlobalHealth_0	83.333	75.302	75	70.024	.014	*
C30_Cognitive_0	83.333	83.454	83.333	82.079	.221	
C30_Sleep_0	33.333	36.983	33.333	37.993	.294	
C30_Emotional_0	75	74.155	75	73.805	.274	

(continued on following page)

TABLE A2. Differences in Features Between Classes in All Target Variables and the *P* Value Generated by Using Mann-Whitney *U* Test for All Features but AgeStudyEntry (*t*-test used), as It Was the Only Normally Distributed Variable (continued)

Input Variable	No Chemotherapy Change (n = 175)		Chemotherapy Change (n = 333)		Significance Test	
	Median	Mean/Value	Median	Mean/Value	<i>P</i>	Significance Level
C30_Physical_0	93.333	86.993	86.667	79.068	<.001	**
C30_Role_0	83.333	76.686	66.667	70.796	.032	*
C30_Social_0	83.333	77.168	83.333	71.954	.029	*
QoLEQ5DMob	1	1.276	1	1.511	<.001	**
QoLEQ5DSelCar	1	1.103	1	1.212	.045	*
QoLEQ5DUsuAct	1	1.603	2	1.858	.002	***
QoLEQ5DPain	2	1.776	2	1.899	.046	*
QoLEQ5DAnxDep	2	1.805	2	1.853	.194	

NOTE. Mean, counts of values for categorical variables, and mean for continuous variables are reported.

**P* ≤ .05.

***P* ≤ .001.

****P* ≤ .01.

TABLE A3. Results of ANOVA Which Provided *P* < .05 for Outcomes, Preprocessing, and Model Comparisons

Factors Affecting Performance	Methods With Significantly Different AUCs
Outcome	Changes to chemotherapy (0.617)—admissions (0.579)*
	Changes to chemotherapy (0.617)—triage (0.543)**
	Admissions (0.579)—triage (0.543)**
Preprocessing	Balanced (0.671)—original (0.536)**
	Balanced (0.671)—partially balanced (0.533)**
Model	DT (0.604)—NN (0.541)***
	SVM (0.615)—NN (0.541)*

Abbreviations: ANOVA, analysis of variance; DT, decision tree; NN, neural network; SVMs, support vector machines.

**P* ≤ .01.

***P* ≤ .001.

****P* ≤ .05.

TABLE A4. Results of the Six Models for Each Target Variable, Each Feature Set, and Each Preprocessing Method Addressing Class Imbalance With Hyperparameters Selected Through Grid Search

		Original Data						Balanced Data						Balanced Train, Real Test							
Model/Outcome	Input Variable	Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters		
LR	Admissions	Clinical variables	0.647	0.667	0.654	0.1	0.513	c = 10, solver = lbfgs	0.654	0.7	0.653	0.676	0.655	c = 10, solver = lblinear	0.569	0.368	0.412	0.389	0.529	c = 1, solver = lblinear	
		All variables	0.738	0.476	0.476	0.476	0.605	c = 100, solver = lbfgs	0.694	0.721	0.721	0.723	0.691	c = 100, solver = newtoncg	0.595	0.406	0.464	0.433	0.563	c = 100, solver = lbfgs	
		PROMs	0.75	0.5	0.238	0.323	0.579	c = 0.01, solver = newtoncg	0.613	0.667	0.59	0.626	0.616	c = 0.1, solver = lbfgs	0.595	0.412	0.5	0.452	0.571	c = 0.01, solver = lbfgs	
		Clinical + C30	0.75	0.5	0.476	0.488	0.659	c = 10, solver = lblinear	0.676	0.727	0.656	0.69	0.676	c = 100, solver = lbfgs	0.607	0.439	0.643	0.522	0.616	c = 1, solver = lblinear	
		Clinical + EQ-SD	0.637	0	0	0	0.5	c = 0.01, solver = newtoncg	0.64	0.703	0.6	0.647	0.644	c = 100, solver = lbfgs	0.539	0.333	0.382	0.356	0.5	c = 1, solver = lblinear	
		Clinical + FACT-G	0.647	0.545	0.162	0.25	0.543	c = 1, solver = lblinear	0.632	0.687	0.613	0.648	0.635	c = 1, solver = lblinear	0.578	0.4	0.529	0.456	0.566	c = 100, solver = newtoncg	
		Triage	Clinical variables	0.588	0.641	0.781	0.704	0.922	c = 100, solver = newtoncg	0.598	0.571	0.6	0.585	0.567	c = 100, solver = lbfgs	0.51	0.567	0.644	0.603	0.485	c = 1, solver = lblinear
		All variables	0.631	0.656	0.824	0.73	0.578	c = 10, solver = newtoncg	0.573	0.491	0.65	0.559	0.584	c = 1, solver = lblinear	0.429	0.5	0.375	0.429	0.438	c = 0.1, solver = newtoncg	
		PROMs	0.643	0.667	0.824	0.737	0.594	c = 10, solver = lblinear	0.594	0.514	0.475	0.494	0.577	c = 0.1, solver = lblinear	0.571	0.625	0.625	0.625	0.563	c = 0.01, solver = lblinear	
		Clinical + C30	0.631	0.667	0.784	0.721	0.589	c = 10, solver = lblinear	0.594	0.51	0.625	0.562	0.598	c = 0.1, solver = lblinear	0.548	0.614	0.563	0.587	0.545	c = 10, solver = lblinear	
		Clinical + EQ-SD	0.578	0.644	0.734	0.686	0.525	c = 10, solver = lbfgs	0.508	0.515	0.583	0.547	0.507	c = 1, solver = lblinear	0.549	0.6	0.661	0.629	0.528	c = 10, solver = newtoncg	
		Clinical + FACT-G	0.598	0.646	0.797	0.713	0.53	c = 10, solver = lbfgs	0.585	0.577	0.683	0.626	0.583	c = 0.1, solver = newtoncg	0.608	0.694	0.576	0.63	0.624	c = 100, solver = lbfgs	
		Chemo	Clinical variables	0.647	0.712	0.776	0.743	0.588	c = 1, solver = lblinear	0.642	0.566	0.684	0.619	0.647	c = 0.1, solver = lblinear	0.627	0.872	0.507	0.642	0.682	c = 0.1, solver = newtoncg
		All variables	0.69	0.75	0.778	0.764	0.656	c = 10, solver = newtoncg	0.732	0.7	0.778	0.737	0.734	c = 1, solver = lblinear	0.69	0.8	0.714	0.755	0.649	c = 10, solver = lblinear	
		PROMs	0.655	0.681	0.87	0.764	0.569	c = 1, solver = lblinear	0.58	0.566	0.556	0.561	0.58	c = 0.01, solver = newtoncg	0.468	0.651	0.5	0.566	0.482	c = 0.1, solver = lbfgs	
		Clinical + C30	0.714	0.742	0.832	0.793	0.659	c = 10, solver = lblinear	0.688	0.656	0.741	0.696	0.689	c = 1, solver = lblinear	0.607	0.767	0.599	0.667	0.616	c = 1, solver = lbfgs	
		Clinical + EQ-SD	0.647	0.653	0.985	0.786	0.493	c = 0.01, solver = lbfgs	0.672	0.594	0.719	0.651	0.678	c = 0.1, solver = lbfgs	0.696	0.821	0.687	0.748	0.7	c = 0.1, solver = newtoncg	
		Clinical + FACT-G	0.608	0.69	0.731	0.71	0.551	c = 1, solver = newtoncg	0.627	0.552	0.649	0.597	0.63	c = 0.1, solver = lblinear	0.598	0.71	0.597	0.661	0.599	c = 0.1, solver = newtoncg	
		OT																			
	Admissions	Clinical variables	0.618	0.25	0.027	0.049	0.49	gini, 2,2,2	0.75	0.773	0.773	0.773	0.747	log_loss, 14,2,4	0.569	0.308	0.235	0.267	0.485	gini, 14,2,2	
			All variables	0.738	0.444	0.19	0.267	0.556	entropy, 2,14,2	0.793	0.779	0.869	0.822	0.784	log_loss, 12,2,4	0.583	0.36	0.321	0.34	0.518	entropy, 12,2,8
		PROMs	0.738	0.444	0.19	0.267	0.556	gini, 2,4,2	0.62	0.847	0.62	0.833	0.82	gini, 18,2,2	0.607	0.4	0.397	0.377	0.546	log_loss, 18,2,4	
		Clinical + C30	0.631	0.321	0.429	0.367	0.563	entropy, 8,18,14	0.62	0.85	0.803	0.831	0.822	gini, 20,2,8	0.619	0.433	0.464	0.448	0.58	gini, 12,4,2	
		Clinical + EQ-SD	0.618	0.464	0.351	0.4	0.56	gini, 20,2,20	0.757	0.733	0.88	0.8	0.743	entropy, 14,2,2	0.588	0.395	0.441	0.417	0.551	entropy, 20,4,2	
		Clinical + FACT-G	0.578	0.412	0.378	0.394	0.535	entropy, 18,4,2	0.779	0.785	0.827	0.805	0.774	gini, 14,2,2	0.578	0.32	0.235	0.271	0.493	gini, 18,2,2	
		Triage	Clinical variables	0.549	0.615	0.75	0.676	0.48	gini, 2,2,2	0.61	0.652	0.6	0.566	0.612	log_loss, 18,2,4	0.51	0.576	0.576	0.576	0.497	entropy, 20,2,4
		All variables	0.583	0.643	0.706	0.673	0.55	gini, 12,18,14	0.698	0.641	0.625	0.633	0.688	gini, 18,2,4	0.5	0.554	0.646	0.596	0.475	entropy, 20,2,2	
		PROMs	0.607	0.688	0.547	0.667	0.596	log_loss, 18,4,2	0.604	0.525	0.525	0.525	0.593	gini, 12,2,2	0.571	0.65	0.542	0.591	0.575	gini, 12,2,2	
		Clinical + C30	0.488	0.611	0.431	0.506	0.504	log_loss, 12,12,5	0.677	0.615	0.6	0.608	0.666	log_loss, 14,2,4	0.524	0.574	0.646	0.608	0.503	log_loss, 18,2,4	
		Clinical + EQ-SD	0.52	0.619	0.609	0.614	0.489	gini, 12,2,18	0.661	0.717	0.55	0.623	0.663	entropy, 20,2,2	0.539	0.603	0.593	0.598	0.529	log_loss, 20,2,2	
		Clinical + FACT-G	0.598	0.635	0.844	0.725	0.514	entropy, 4,2,4	0.669	0.691	0.633	0.661	0.67	entropy, 20,2,2	0.598	0.645	0.678	0.661	0.583	gini, 18,2,4	
		Chemo	Clinical variables	0.657	0.742	0.731	0.737	0.623	log_loss, 4,14,4	0.731	0.698	0.649	0.673	0.721	log_loss, 12,4,14	0.48	0.625	0.522	0.569	0.461	gini, 20,2,12
		All variables	0.655	0.671	0.907	0.772	0.554	gini, 2,8,2	0.75	0.842	0.933	0.696	0.745	log_loss, 18,2,2	0.702	0.763	0.804	0.783	0.652	gini, 20,2,2	
		PROMs	0.571	0.629	0.815	0.71	0.474	entropy, 12,4,20	0.786	0.875	0.648	0.748	0.781	entropy, 20,2,2	0.571	0.656	0.75	0.7	0.482	gini, 14,2,4	
		Clinical + C30	0.655	0.671	0.907	0.772	0.554	entropy, 2,8,2	0.813	0.923	0.667	0.774	0.807	log_loss, 14,4,2	0.583	0.691	0.679	0.685	0.536	entropy, 20,2,4	
		Clinical + EQ-SD	0.667	0.732	0.776	0.754	0.617	entropy, 4,18,2	0.791	0.809	0.667	0.731	0.775	gini, 12,2,2	0.657	0.75	0.716	0.733	0.63	entropy, 20,2,2	
		Clinical + FACT-G	0.647	0.707	0.791	0.746	0.581	log_loss, 4,18,2	0.784	0.759	0.719	0.739	0.775	entropy, 14,4,2	0.569	0.689	0.627	0.656	0.542	log_loss, 18,2,4	
		AB																			
	Admissions	Clinical variables	0.618	0.25	0.027	0.049	0.49	SAMME, 1,0, est = 10	0.566	0.629	0.52	0.569	0.571	SAMME, 1,0, est = 500	0.5	0.257	0.265	0.261	0.441	SAMME, 1,0, est = 500	
			All variables	0.762	0.667	0.095	0.167	0.54	SAMME, 0,01, est = 500	0.676	0.687	0.754	0.719	0.667	SAMME, 1,0, est = 100	0.643	0.469	0.536	0.5	0.616	SAMME, 1,0, est = 500
		PROMs	0.738	0.333	0.048	0.083	0.508	SAMME, 0,01, est = 500	0.649	0.683	0.672	0.678	0.646	SAMME, 1,0, est = 500	0.548	0.393	0.429	0.387	0.518	SAMME, 1,0, est = 500	
		Clinical + C30	0.655	0.3	0.286	0.293	0.532	SAMME, 1,0, est = 50	0.739	0.767	0.754	0.76	0.737	SAMME, 0,1, est = 500	0.595	0.477	0.536	0.469	0.55	SAMME, 1, est = 500	
		Clinical + EQ-SD	0.618	0.25	0.027	0.049	0.49	SAMME, 1,0, est = 50	0.61	0.683	0.547	0.607	0.618	SAMME, 1,0, est = 500	0.461	0.216	0.238	0.225	0.404	SAMME, 1,0, est = 500	
		Clinical + FACT-G	0.627	0	0	0	0.492	SAMME, 0,01, est = 100	0.684	0.716	0.707	0.711	0.687	SAMME, 1,0, est = 500	0.48	0.229	0.235	0.232	0.419	SAMME, 1,0, est = 500	
		Triage	Clinical variables	0.559	0.622	0.797	0.699	0.491	SAMME, 0,01, est = 500	0.568	0.585	0.517	0.549	0.559	SAMME, 1,0, est = 500	0.51	0.579	0.559	0.559	0.501	SAMME, 0,1, est = 100
		All variables	0.512	0.586	0.667	0.624	0.47	SAMME, 0,1, est = 500	0.667	0.583	0.7	0.636	0.671	SAMME, 1,0, est = 500	0.476	0.543	0.521	0.532	0.469	SAMME, 1,0, est = 500	
		PROMs	0.524	0.604	0.627	0.615	0.495	SAMME, 1,0, est = 500	0.646	0.56	0.7	0.622	0.654	SAMME, 1,0, est = 500	0.464	0.532	0.521	0.526	0.455	SAMME, 1,0, est = 500	
		Clinical + C30	0.607	0.61	0.98	0.792	0.595	SAMME, 0,1, est = 10	0.646	0.571	0.6	0.585	0.639	SAMME, 1,0, est = 10	0.524	0.583	0.583	0.583	0.514	SAMME, 1,0, est = 500	
		Clinical + EQ-SD	0.578	0.63	0.797	0.703	0.504	SAMME, 0,01, est = 100	0.525	0.536	0.5	0.517	0.526	SAMME, 1,0, est = 500	0.578	0.618	0.712	0.661	0.554	SAMME, 0,1, est = 500	
		Clinical + FACT-G	0.578	0.63	0.797	0.703	0.504	SAMME, 1,0, est = 10	0.585	0.59	0.6	0.595	0.584	SAMME, 1,0, est = 500	0.49	0.557	0.576	0.567	0.474	SAMME, 1,0, est = 500	

(continued on following page)

TABLE A4. Results of the Six Models for Each Target Variable, Each Feature Set, and Each Preprocessing Method Addressing Class Imbalance With Hyperparameters Selected Through Grid Search (continued)

Model/Outcome	Input Variable	Original Data					Balanced Data					Balanced Train, Real Test							
		Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters
Chemo	Clinical variables	0.647	0.763	0.672	0.714	0.636	SAMME, 0.01, est = 500	0.664	0.581	0.754	0.646	0.676	SAMME, 1.0, est = 500	0.578	0.722	0.592	0.645	0.577	SAMME, 1.0, est = 50
	All variables	0.643	0.643	1	0.783	0.5	SAMME, 0.1, est = 10	0.705	0.684	0.722	0.703	0.706	SAMME, 1, est = 500	0.69	0.768	0.768	0.768	0.652	SAMME, 1.0, est = 500
	PROMs	0.593	0.667	0.704	0.685	0.535	SAMME, 1.0, est = 50	0.652	0.636	0.648	0.642	0.652	SAMME, 1.0, est = 100	0.524	0.643	0.643	0.643	0.464	SAMME, 0.1, est = 500
	Clinical + C30	0.643	0.643	1	0.783	0.5	SAMME, 0.1, est = 100	0.741	0.766	0.667	0.713	0.739	SAMME, 1.0, est = 500	0.524	0.674	0.554	0.608	0.509	SAMME, 1.0, est = 500
	Clinical + EQ-SD	0.657	0.657	1	0.793	0.5	SAMME, 0.1, est = 50	0.701	0.623	0.754	0.683	0.708	SAMME, 1.0, est = 100	0.647	0.754	0.687	0.719	0.629	SAMME, 1.0, est = 500
	Clinical + FACT-G	0.706	0.747	0.836	0.789	0.646	SAMME, 1, est = 100	0.709	0.632	0.754	0.688	0.715	SAMME, 0.1, est = 500	0.657	0.767	0.687	0.724	0.643	SAMME, 1.0, est = 100
RF																			
Admissions	Clinical variables	0.637	0	0	0	0.5	entropy, 2.2, est = 100	0.816	0.813	0.867	0.809	0.81	gini, 142, est = 100	0.471	0.25	0.294	0.27	0.426	gini, 142, est = 100
	All variables	0.702	0.333	0.19	0.242	0.532	gini, 2.8, est = 10	0.901	0.891	0.934	0.912	0.897	gini, 142, est = 500	0.69	0.583	0.25	0.35	0.58	log_loss, 142, est = 100
	PROMs	0.726	0.375	0.143	0.207	0.532	log_loss, 18.4, est = 10	0.865	0.838	0.934	0.884	0.857	gini, 182, est = 100	0.655	0.478	0.393	0.431	0.589	entropy, 182, est = 100
	Clinical + C30	0.762	0.545	0.286	0.375	0.603	entropy, 142, est = 100	0.91	0.892	0.951	0.921	0.905	gini, 142, est = 100	0.655	0.478	0.393	0.431	0.589	gini, 182, est = 100
	Clinical + EQ-SD	0.627	0.495	0.136	0.208	0.521	log_loss, 8.18, est = 10	0.794	0.783	0.867	0.823	0.786	entropy, 182, est = 500	0.539	0.276	0.235	0.254	0.463	gini, 182, est = 500
	Clinical + FACT-G	0.627	0.471	0.216	0.296	0.539	gini, 18.14, est = 100	0.868	0.87	0.893	0.882	0.865	log_loss, 182, est = 100	0.667	0.5	0.206	0.262	0.551	gini, 182, est = 100
Triage	Clinical variables	0.588	0.625	0.859	0.724	0.495	log_loss, 2.18, est = 500	0.61	0.63	0.567	0.597	0.611	entropy, 182, est = 500	0.471	0.536	0.627	0.578	0.441	entropy, 182, est = 500
	All variables	0.512	0.596	0.608	0.602	0.489	entropy, 18.18, est = 10	0.719	0.633	0.775	0.697	0.727	gini, 182, est = 100	0.5	0.552	0.667	0.604	0.472	gini, 182, est = 500
	PROMs	0.607	0.615	0.941	0.744	0.516	gini, 2.18, est = 10	0.677	0.592	0.725	0.652	0.684	log_loss, 182, est = 500	0.5	0.545	0.75	0.632	0.458	entropy, 142, est = 500
	Clinical + C30	0.536	0.604	0.686	0.642	0.495	log_loss, 5.14, est = 10	0.696	0.604	0.8	0.688	0.713	gini, 182, est = 500	0.5	0.55	0.667	0.604	0.472	gini, 182, est = 100
	Clinical + EQ-SD	0.598	0.635	0.844	0.725	0.514	entropy, 2.14, est = 10	0.703	0.705	0.717	0.711	0.703	gini, 182, est = 100	0.618	0.645	0.647	0.719	0.575	entropy, 142, est = 100
	Clinical + FACT-G	0.608	0.633	0.891	0.74	0.511	entropy, 2.18, est = 10	0.686	0.695	0.683	0.689	0.686	log_loss, 182, est = 100	0.588	0.605	0.831	0.7	0.543	gini, 142, est = 500
Chemo	Clinical variables	0.657	0.739	0.701	0.75	0.623	log_loss, 5.18, est = 100	0.799	0.778	0.737	0.757	0.79	log_loss, 142, est = 500	0.637	0.734	0.701	0.718	0.658	gini, 182, est = 500
	All variables	0.643	0.643	1	0.783	0.5	log_loss, 2.18, est = 10	0.902	0.906	0.889	0.897	0.902	entropy, 142, est = 500	0.774	0.776	0.929	0.846	0.696	gini, 142, est = 500
	PROMs	0.643	0.643	1	0.783	0.5	log_loss, 2.18, est = 10	0.804	0.864	0.704	0.776	0.8	log_loss, 142, est = 100	0.649	0.681	0.875	0.786	0.527	log_loss, 182, est = 500
	Clinical + C30	0.643	0.643	1	0.783	0.5	entropy, 2.14, est = 10	0.902	0.906	0.889	0.897	0.901	entropy, 182, est = 500	0.619	0.722	0.696	0.709	0.58	entropy, 182, est = 500
	Clinical + EQ-SD	0.686	0.727	0.836	0.778	0.618	gini, 14.18, est = 100	0.881	0.887	0.825	0.895	0.873	entropy, 182, est = 500	0.676	0.75	0.806	0.766	0.617	gini, 182, est = 100
	Clinical + FACT-G	0.627	0.704	0.756	0.725	0.573	entropy, 14.14, est = 10	0.858	0.88	0.772	0.822	0.847	log_loss, 142, est = 100	0.667	0.72	0.806	0.761	0.603	gini, 182, est = 500
BNM																			
Admissions	Clinical variables	0.637	0	0	0	0.5	c = 0.01, gamma = 0.001, kernel = linear	0.75	0.766	0.787	0.775	0.746	c = 1, gamma = 1, kernel = rbf	0.40	0.2	0.176	0.188	0.411	c = 1, gamma = 1, kernel = rbf
	All variables	0.738	0.462	0.286	0.353	0.587	c = 0.1, gamma = 0.001, kernel = linear	0.892	1	0.803	0.891	0.902	c = 1, gamma = 0.1, kernel = rbf	0.667	0	0	0	0.5	c = 0.1, gamma = 1, kernel = rbf
	PROMs	0.726	0.333	0.095	0.148	0.516	c = 1, gamma = 0.01, kernel = rbf	0.45	0	0	0	0.5	c = 0.1, gamma = 1, kernel = rbf	0.667	0.5	0.036	0.087	0.509	c = 1, gamma = 1, kernel = rbf
	Clinical + C30	0.75	0.5	0.286	0.364	0.595	c = 1, gamma = 0.001, kernel = linear	0.874	1	0.77	0.87	0.885	c = 1, gamma = 0.01, kernel = rbf	0.667	0	0	0	0.5	c = 1, gamma = 1, kernel = rbf
	Clinical + EQ-SD	0.637	0	0	0	0.5	c = 0.01, gamma = 0.01, kernel = rbf	0.787	0.859	0.733	0.791	0.793	c = 1, gamma = 1, kernel = rbf	0.618	0.273	0.088	0.133	0.485	c = 1, gamma = 1, kernel = rbf
	Clinical + FACT-G	0.637	0	0	0	0.5	c = 0.1, gamma = 0.001, kernel = poly	0.838	0.896	0.8	0.845	0.843	c = 1, gamma = 1, kernel = rbf	0.647	0.429	0.176	0.25	0.529	c = 1, gamma = 1, kernel = rbf
Triage	Clinical variables	0.559	0.62	0.766	0.685	0.488	c = 1, gamma = 0.01, kernel = linear	0.559	0.554	0.683	0.612	0.557	c = 1, gamma = 0.1, kernel = rbf	0.49	0.554	0.61	0.581	0.468	c = 1, gamma = 1, kernel = rbf
	All variables	0.643	0.662	0.843	0.741	0.588	c = 1, gamma = 0.001, kernel = linear	0.729	0.609	0.975	0.75	0.764	c = 1, gamma = 1, kernel = rbf	0.583	0.578	1	0.733	0.514	c = 1, gamma = 1, kernel = rbf
	PROMs	0.619	0.642	0.843	0.729	0.558	c = 1, gamma = 0.001, kernel = linear	0.417	0.417	1	0.588	0.5	c = 0.1, gamma = 1, kernel = rbf	0.583	0.58	0.979	0.729	0.517	c = 0.1, gamma = 1, kernel = rbf
	Clinical + C30	0.607	0.607	1	0.756	0.5	c = 0.1, gamma = 0.001, kernel = sigmoid	0.688	0.589	0.825	0.686	0.707	c = 1, gamma = 0.1, kernel = rbf	0.56	0.566	0.979	0.718	0.49	c = 0.01, gamma = 1, kernel = rbf
	Clinical + EQ-SD	0.549	0.618	0.734	0.671	0.486	c = 0.01, gamma = 0.01, kernel = poly	0.763	0.729	0.85	0.785	0.781	c = 1, gamma = 1, kernel = rbf	0.578	0.593	0.854	0.703	0.525	c = 1, gamma = 1, kernel = rbf
	Clinical + FACT-G	0.569	0.632	0.75	0.686	0.507	c = 0.1, gamma = 0.01, kernel = poly	0.746	0.688	0.917	0.786	0.743	c = 1, gamma = 1, kernel = rbf	0.578	0.585	0.932	0.719	0.513	c = 1, gamma = 1, kernel = rbf
Chemo	Clinical variables	0.657	0.735	0.746	0.741	0.616	c = 1, gamma = 0.1, kernel = rbf	0.761	0.712	0.737	0.724	0.758	c = 1, gamma = 1, kernel = rbf	0.637	0.8	0.597	0.68	0.656	c = 1, gamma = 1, kernel = rbf
	All variables	0.643	0.643	1	0.783	0.5	c = 0.01, gamma = 0.001, kernel = linear	0.929	0.871	1	0.931	0.931	c = 1, gamma = 1, kernel = rbf	0.667	0.667	1	0.8	0.5	c = 1, gamma = 1, kernel = rbf
	PROMs	0.631	0.695	0.759	0.726	0.58	c = 1, gamma = 0.001, kernel = poly	0.902	0.841	0.981	0.905	0.905	c = 1, gamma = 1, kernel = rbf	0.679	0.679	0.982	0.803	0.527	c = 1, gamma = 1, kernel = rbf
	Clinical + C30	0.762	0.793	0.852	0.821	0.726	c = 1, gamma = 0.001, kernel = linear	0.929	0.871	1	0.931	0.931	c = 1, gamma = 1, kernel = rbf	0.655	0.663	0.982	0.791	0.491	c = 1, gamma = 1, kernel = rbf
	Clinical + EQ-SD	0.657	0.742	0.687	0.713	0.615	c = 0.1, gamma = 0.001, kernel = poly	0.981	0.847	0.877	0.862	0.88	c = 1, gamma = 1, kernel = rbf	0.657	0.7	0.836	0.762	0.575	c = 1, gamma = 1, kernel = rbf
	Clinical + FACT-G	0.627	0.71	0.731	0.721	0.58	c = 0.1, gamma = 0.01, kernel = poly	0.886	0.82	0.877	0.847	0.867	c = 1, gamma = 1, kernel = rbf	0.676	0.693	0.91	0.787	0.57	c = 1, gamma = 1, kernel = rbf
NN																			
Admissions	Clinical variables	0.637	0	0	0	0.5	relu, 0.0001, (150,50,30), adam	0.493	0.615	0.213	0.317	0.525	tanh, 0.0001, (100,50,20), adam	0.333	0.33	0.97	0.493	0.493	tanh, 0.0001, (150,50,30), adam
	All variables	0.726	0.4	0.19	0.258	0.548	tanh, 0.001, (100,50,20), adam	0.73	0.763	0.738	0.75	0.729	tanh, 0.0001, (150,50,30), adam	0.643	0.438	0.25	0.318	0.545	tanh, 0.001, (150,50,30), adam
	PROMs	0.667	0.333	0.333	0.333	0.556	relu, 0.01, (100), adam	0.685	0.717	0.705	0.711	0.682	tanh, 0.0001, (150,50,30), adam	0.548	0.273	0.214	0.24	0.464	tanh, 0.001, (150,50,30), adam
	Clinical + C30	0.702	0.35	0.095	0.138	0.5	tanh, 0.01, (100,50,20), adam	0.766	0.787	0.787	0.787	0.763	tanh, 0.0001, (150,50,30), adam	0.607	0.407	0.393	0.4	0.554	tanh, 0.001, (100,50,20), adam
	Clinical + EQ-SD	0.637	0	0	0	0.5	relu, 0.001, (100), sgdr	0.625	0.607	0.907	0.727	0.933	tanh, 0.001, (150,50,30), adam	0.62	0.347	0.5	0.43	0.515	tanh, 0.001, (150,50,30), adam
	Clinical + FACT-G	0.647	1	0.027	0.083	0.514	tanh, 0.0001, (100,50,20), adam	0.537	0.7	0.28	0.4	0.566	tanh, 0.01, (150,50,30), adam	0.533	0.333	1	0.5	0.5	tanh, 0.01, (100,50,20), adam
(continued on following page)																			

TABLE A4. Results of the Six Models for Each Target Variable, Each Feature Set, and Each Preprocessing Method Addressing Class Imbalance With Hyperparameters Selected Through Grid Search (continued)

Model/Outcome	Input Variable	Original Data						Balanced Data						Balanced Train, Real Test					
		Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters	Accuracy	Precision	Recall	F1	AUC	Hyperparameters
Triage	Clinical variables	0.627	0.627	1	0.771	0.5	tanh, 0.001, (150,50,20), sgd	0.492	0	0	0	0.5	tanh, 0.001, (100,50,20), adam	0.569	0.579	0.622	0.714	0.601	tanh, 0.0001, (150,60,30), adam
	All variables	0.631	0.628	0.961	0.76	0.541	relu, 0.0001, (100), adam	0.573	0.486	0.425	0.435	0.552	tanh, 0.0001, (150,60,30), adam	0.464	0.543	0.296	0.459	0.476	tanh, 0.001, (150,60,30), adam
	PROMs	0.548	0.61	0.705	0.655	0.504	relu, 0.01, (150,60,30), sgd	0.563	0.471	0.4	0.432	0.539	relu, 0.0001, (150,60,30), adam	0.536	0.6	0.563	0.581	0.531	tanh, 0.00001, (150,60,30), adam
	Clinical + C30	0.56	0.613	0.745	0.673	0.509	relu, 0.01, (150,60,30), sgd	0.427	0.405	0.8	0.538	0.48	tanh, 0.01, (150,60,30), adam	0.524	0.574	0.646	0.608	0.503	tanh, 0.001, (150,60,30), adam
	Clinical + EQ-5D	0.618	0.624	0.984	0.764	0.492	relu, 0.0001, (150,60,30), adam	0.559	0.6	0.4	0.48	0.56	tanh, 0.01, (150,60,30), adam	0.598	0.598	0.932	0.738	0.536	tanh, 0.001, (150,60,30), adam
	Clinical + FACT-G	0.627	0.627	1	0.771	0.5	relu, 0.0001, (100), sgd	0.483	0.495	0.883	0.635	0.476	tanh, 0.0001, (100), adam	0.588	0.584	1	0.737	0.512	tanh, 0.0001, (100,50,20), sgd
Chemo	Clinical variables	0.637	0.65	0.97	0.778	0.485	relu, 0.01, (100,50,20), sgd	0.425	0.425	1	0.597	0.5	tanh, 0.01, (150,60,30), adam	0.373	0.8	0.06	0.111	0.516	tanh, 0.001, (100), sgd
	All variables	0.607	0.648	0.852	0.736	0.509	relu, 0.0001, (150,60,30), sgd	0.607	0.594	0.799	0.651	0.612	tanh, 0.001, (150,60,30), adam	0.619	0.662	0.875	0.754	0.491	tanh, 0.001, (150,60,30), adam
	PROMs	0.607	0.64	0.889	0.744	0.494	tanh, 0.001, (100), adam	0.616	0.622	0.519	0.595	0.613	tanh, 0.001, (150,60,30), adam	0.643	0.741	0.716	0.727	0.607	tanh, 0.001, (150,60,30), adam
	Clinical + C30	0.643	0.643	1	0.783	0.5	relu, 0.001, (100), sgd	0.571	0.583	0.389	0.467	0.565	tanh, 0.01, (150,60,30), adam	0.607	0.702	0.714	0.708	0.554	tanh, 0.001, (150,60,30), adam
	Clinical + EQ-5D	0.657	0.657	1	0.793	0.5	relu, 0.00001, (100,50,20), sgd	0.634	0.551	0.754	0.637	0.65	tanh, 0.0001, (150,60,30), adam	0.676	0.815	0.687	0.727	0.686	tanh, 0.0001, (150,60,30), adam
	Clinical + FACT-G	0.657	0.674	0.925	0.78	0.534	tanh, 0.001, (100,50,20), adam	0.53	0.471	0.842	0.604	0.57	tanh, 0.00001, (150,60,30), adam	0.51	0.73	0.403	0.519	0.559	tanh, 0.001, (150,60,30), adam

NOTE. Hyperparameters considered for LR were regularization strength and solver. Hyperparameters considered for DT were criterion, maximum depth, the minimum number samples required to split, and minimum number of samples required to be at a leaf node. Hyperparameters considered for AB were number of estimators, learning rate, and boosting algorithm. Hyperparameters considered for RF were criterion, maximum depth, the minimum number samples required to split, and the number of estimators. Hyperparameters considered for SVM were kernel type, kernel coefficient, and regularization parameter. Hyperparameters considered for NN were activation function for the hidden layer, solver for weight optimization, the number of neurons in hidden layers, and the strength of the L2 regularization term.

Abbreviations: AB, adaptive boosting; DT, decision tree; EQ-5D, EuroQol Five-Dimensional Visual Analogue Scale; FACT-G, Functional Assessment of Cancer Therapy-General; LR, logistic regression; NN, neural network; PROMs, patient-reported outcome measures; QLQ-C30, European Organisation for Research and Treatment of Cancer Core Quality of Life Questionnaire; RF, random forest; SVM, support vector machine. Highest values achieved by all models in all target variables are set in bold.

8.3 Paper 3

Machine Learning Models Predicting Hospital Admissions During Chemotherapy Utilising Longitudinal Symptom Severity Reports and Patient-Reported Outcome Measures

Zuzanna WÓJCIK^{a,b,1}, Vania DIMITROVA^b, Lorraine WARRINGTON^c, Galina VELIKOVA^{c,d}, Kate ABSOLOM^{c,e}, and Samuel D. RELTON^e

^a *UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care, University of Leeds, Leeds, UK* ^b *School of Computer Science, University of Leeds, UK*

^c *Leeds Institute of Medical Research, St James's University Hospital, Leeds, UK* ^d *Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, Leeds, UK* ^e *Leeds Institute of Health Sciences, Leeds, UK*

ORCID ID: Zuzanna Wójcik <https://orcid.org/0009-0007-6214-7736>, Vania Dimitrova <https://orcid.org/0000-0002-7001-0891>, Lorraine Warrington <https://orcid.org/0000-0002-8389-6134>, Galina Velikova <https://orcid.org/0000-0003-1899-5942>, Kate Absolom <https://orcid.org/0000-0002-5477-6643>, Samuel D. Relton <https://orcid.org/0000-0003-0634-4587>

Abstract. Chemotherapy toxicity can lead to acute hospital admissions, negatively impacting the healthcare system and patients' well-being. Machine learning (ML) models identifying patients at risk of emergency admissions are often developed on data lacking patients' perspective. This study used longitudinally collected symptom severity reports and 4 ML models to predict hospital admissions risk during chemotherapy, and short-term admissions risk (within 14 days of a report). It also compared performance of models developed with, and without the use of patient-reported outcome measures (PROMs). Random forest and extreme gradient boosting models predicted admissions with excellent balanced accuracy, recall, and specificity of over 0.9. However, short-term admissions risk predictions were poor. PROMs improved overall model performance. The results advocate for longitudinal collection and use of symptom severity reports and PROMs. This can support understanding of chemotherapy toxicity patterns leading to emergency admissions, and inform clinicians and patients of potential future complications.

Keywords. Patient-reported data, Hospital admissions predictions, ML

¹ Corresponding Author: Zuzanna Wójcik; E-mail: sczw@leeds.ac.uk.

1. Introduction

Emergency hospital admissions resulting from chemotherapy toxicity can negatively impact patients' quality of life (QoL) and the healthcare system. Machine learning (ML) has been successful in predicting acute hospitalisation during cancer treatment [1], which can help to identify patients at risk of severe chemotherapy toxicity, plan for emergency admissions, and inform treatment decisions.

Nevertheless, studies predicting clinical outcomes often neglect patients' perspective and develop models using only clinical and demographic data. Our previous work has shown that including patient-reported outcome measures (PROMs), which capture patients' QoL, improves model performance predicting emergency admissions [2]. However, the data were collected only at chemotherapy baseline.

Symptom severity reports are also patient-reported data, often collected longitudinally. They provide patients' perspective on their health status throughout the duration of treatment. The patterns in longitudinal reports can be identified by ML models to predict long-term, and short-term clinical outcomes [3]. However, limited research has been conducted on predictive value of longitudinal symptom severity data in predicting hospital utilisation.

This study aims to predict the risk of hospital admissions occurring during chemotherapy, and the short-term admissions risk (within 14 days from a report), using 4 ML models and longitudinal symptom severity reports, clinical, and demographic data. To establish the PROMs' predictive value, we compared performances of models including and excluding PROMs from feature sets.

2. Methods

Overall methodology: Logistic regression (LR), random forest (RF), multilayer perceptron (MLP), and extreme gradient boosting (XGB) were developed using symptom severity reports, clinical, demographic, and PROMs (included, or excluded) data to predict: 1) admissions during chemotherapy at any time after a report; 2) admissions risk within 14 days of a report (Figure 1). The models were selected due to their common use in similar studies [4]. A clinical oncologist was involved in the study design to ensure relevance of this study to clinical practice.

Dataset: The data (3260 records) were collected from 254 patients initiating chemotherapy for colorectal, breast, or gynecological cancer during eRAPID randomised clinical trial [5]. The patients were asked to complete symptom severity reports weekly for 18 weeks. Irregularity of reports were coded by a feature representing time since previous report. PROMs (obtained from Five-dimensional Visual Analogue Scale (EQ-5D-VAS), Functional Assessment of Cancer Therapy - General 28 items (FACT-G), and EORTC Core Quality of Life Questionnaire (QLQ-C30), cited in Absolom et al., 2021 [5]) were completed at baseline, 6, 12, and 18 weeks (most recent value selected for each record). Clinical and demographic data were derived from electronic healthcare records. The target variables - admissions during chemotherapy (following a report) and admissions within 14 days from a report were computed from clinical event dates. The

14 days window was chosen with input from oncology expertise to reflect a useful time frame for detecting change in clinical practice and alerting of potential admission.

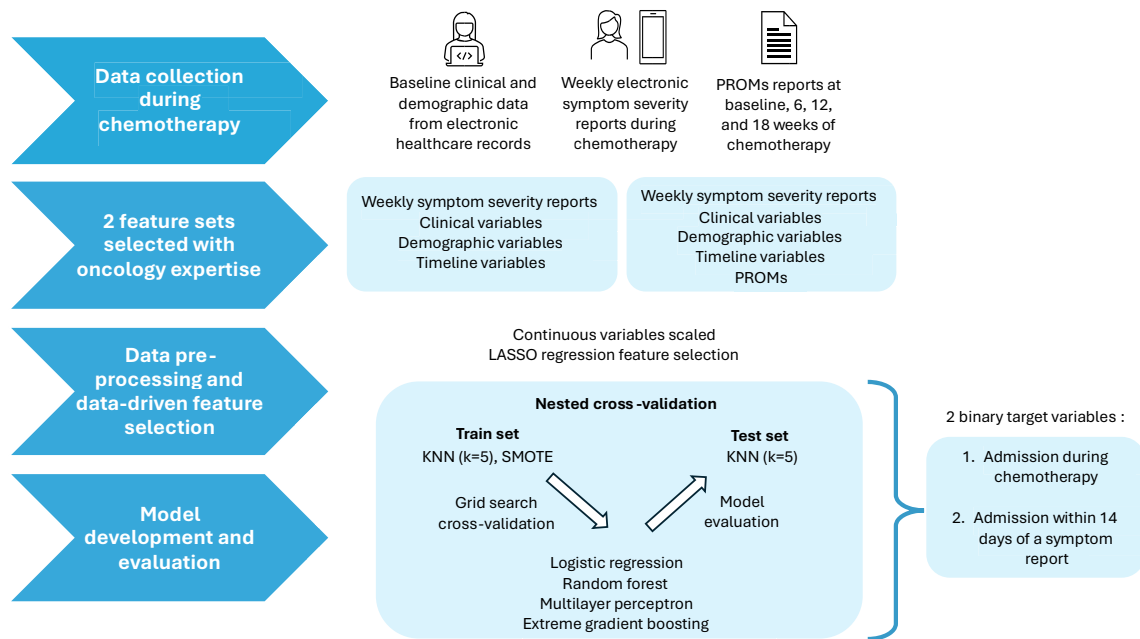


Figure 1. Flow diagram illustrating study methodology.

Data pre-processing: Details of data pre-processing are presented in Figure 1. Due to a class imbalance (640 reports followed by an admission after report, and 175 reports within 14 days), Synthetic Minority Oversampling Technique (SMOTE) was applied on train set (80%) to balance the data. Test set (20%) was left imbalanced, enabling model evaluation on real patients' data and relevance to clinical practice. Feature selection was consulted with oncologist and further conducted using Least Absolute Shrinkage and Selection Operator (LASSO) regression with cross-validated minimised lambda value to reduce bias [6].

Model development and evaluation: Nested cross-validation was performed during model development and evaluation to prevent data leakage. LR, RF, MLP, and XGB were developed with grid search hyperparameter tuning, performed on train set. Models with hyperparameters maximising balanced accuracy (BA) were evaluated on test set. BA, specificity (SP), and recall (TPR) were the main performance metrics, as they best evaluate models as screening systems, providing information on correct classification for both (even imbalanced) classes. Accuracy, precision, F1 score, and area under the ROC curve (AUC) were also reported. All model hyperparameters, results, and selected features are provided in supplementary file: <https://github.com/zuzawojcik/MIE>.

3. Results

Admissions predictions: Admissions during chemotherapy predictions had higher performance than admissions within 14 days of completed reports. All models achieved an outstanding BA of over 0.9. These models also had very high SP and TPR, suggesting

successful predictions. In contrast, predicting admissions within the 14-day interval resulted in poor BAs (below 0.7). The SPs were very high for all models, apart from LR, but TPRs were extremely compromised, with the lowest values being below 0.2 (Figure 2).

Model comparison: In predicting admissions during chemotherapy, XGB had the highest BA (0.989) and all other metrics, indicating a great overall performance. MLP classifier was the least successful, although all its metrics were above 0.8. In contrast, BA of predicting admissions within 14 days was the highest for MLP, and the lowest for XGB. SPs of all models, except for LR were high, but TPRs were compromised.

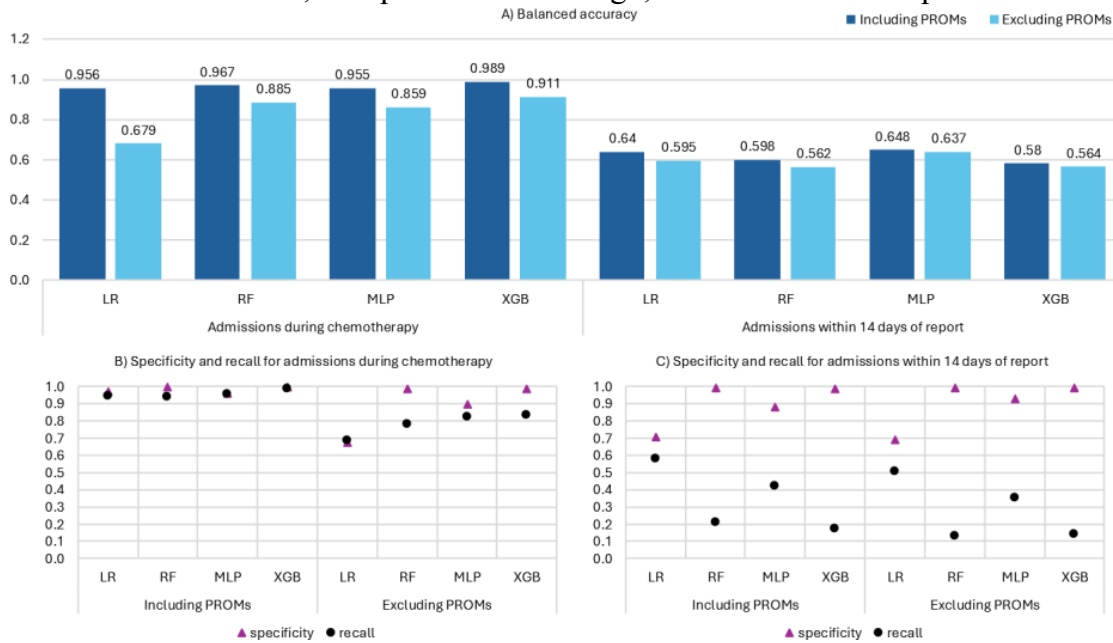


Figure 2. Balanced accuracy, specificity, and recall of all models.

PROMs as input features: As expected, overall performance of models increased when PROMs variables were present. All models in all scenarios had higher BA when PROMs were included in analysis. PROMs especially increased the performance of LR model predicting admissions during chemotherapy. Overall, including PROMs also improved TPR values for admissions within 14 days interval predictions, but SP values were mostly unaffected.

4. Discussion

Overall findings: This study successfully predicted hospital admissions during chemotherapy from longitudinally collected symptom severity reports and PROMs data using ML. The models exceeded performance of admissions predictions from the study using baseline data only [2], encouraging the collection and use of symptom severity reports. The potential explanation for poor performance of predicting admissions within 14-day interval could be the inherent class imbalance, which was increased due to the shorter time-frame of interest. As in previous studies [2], PROMs improved model performance, which emphasised the high value of patients' subjective view on their health status.

Potential applications and strengths: Identifying patients at risk of acute admissions has the potential to forewarn both emergency healthcare workers and patients. The awareness chemotherapy toxicity likelihood could support shared decision making prior to, and during chemotherapy. Using longitudinal patient reported data reveals the patients' perspective on their health and increases the patients' participation in symptom management.

Future work and limitations: The inevitable consequence of using clinical trial data is the limited representability of study population. Therefore, external validation of these models is a crucial next step to ensure their real-world application. Longitudinal reports provide an opportunity for more complex modeling using deep learning methods, such as recurrent neural networks [3]. These could improve the more immediate admissions predictions following the reports.

5. Conclusions

This study incorporated longitudinal symptom severity reports and PROMs in ML models to predict emergency hospital admissions during chemotherapy, and admissions within 14 days of report. Overall results indicated excellent performance of models predicting admissions during chemotherapy, which advocates for patient-reported data collection. This has the potential to inform cancer treatment decisions and support planning for emergency hospitalisations.

Acknowledgements. The authors thank patients and clinicians from eRAPID clinical trial for providing the data. Ethics approval was granted by School of Medicine Research Ethics Committee [MREC 22-098], University of Leeds. This work was supported in part by UK Research and Innovation (UKRI) [EP/S024336/1].

References

- [1] Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm. *JCO Clinical Cancer Informatics*. 2018 Dec;(2):1-11.
- [2] Wójcik Z, Dimitrova V, Warrington L, Velikova G, Absolom K. Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From Patient Reported Outcome Measures. *JCO Clinical Cancer Informatics*. 2024 Apr;(8):e2300264. Publisher: Wolters Kluwer.
- [3] Wang Y, Van Dijk L, Mohamed ASR, Fuller CD, Zhang X, Marai GE, et al. Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes. *Proceedings International Database Engineering and Applications Symposium*. 2021 Jul;2021:273-9.
- [4] Krepper D, Cesari M, Hubel NJ, Zelger P, Sztankay MJ. Machine learning models including patient-reported outcome data in oncology: a systematic literature review and analysis of their reporting quality. *Journal of Patient-Reported Outcomes*. 2024 Nov;8(1):126.
- [5] Absolom K, Warrington L, Hudson E, Hewison J, Morris C, Holch P, et al. Phase III Randomized Controlled Trial of eRAPID: eHealth Intervention During Chemotherapy. *Journal of Clinical Oncology*. 2021 Mar;39(7):734-47. Publisher: Wolters Kluwer.
- [6] Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. In: 2016 IEEE International Conference on Advances in Computer Applications (ICACA); 2016. p. 18-20.

8.4 Paper 4



UNIVERSITY OF LEEDS

This is a repository copy of *Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes During Chemotherapy*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/215823/>

Version: Accepted Version

Proceedings Paper:

Wójcik, Z., Dimitrova, V. orcid.org/0000-0002-7001-0891, Warrington, L. orcid.org/0000-0002-8389-6134 et al. (2 more authors) (2024) Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes During Chemotherapy. In: Finkelstein, J., Moskovitch, R. and Parimbelli, E., (eds.) Artificial Intelligence in Medicine. 22nd International Conference, AIME 2024, 09-12 Jul 2024, Salt Lake City, UT, USA. Lecture Notes in Computer Science, 14844 . Springer Nature , pp. 101-116. ISBN: 978-3-031-66538-7 ISSN: 0302-9743 EISSN: 1611-3349

https://doi.org/10.1007/978-3-031-66538-7_12

© The Authors 2024. This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use (<https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>), but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-031-66538-7_12

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Patient-Centric Approach for Utilising Machine Learning to Predict Health-Related Quality of Life Changes during Chemotherapy

Zuzanna Wójcik^{1,2}[0009-0007-6214-7736], Vania Dimitrova²[0000-0002-7001-0891],
Lorraine Warrington³[0000-0002-8389-6134], Galina
Velikova^{3,4}[0000-0003-1899-5942], and Kate Absolom^{3,5}[0000-0002-5477-6643]

¹ UKRI Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care, Leeds, UK

² School of Computing, University of Leeds, Leeds, United Kingdom

³ Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, Leeds, UK

⁴ Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, Leeds, UK

⁵ Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

Abstract. Patients undergoing chemotherapy often experience adverse effects, which can lead to changes in health-related quality of life (HRQOL) and have detrimental effects on patients' physical and psychological well-being. This study aims to apply machine learning (ML) models to patient-reported, clinical, and demographic data to predict changes in physical well-being, social functioning, role functioning, usual activities, and mobility at 6, 12 and 18 weeks from starting chemotherapy. A patient-centric approach is followed as outcome variables were selected after consultation with patients and a clinician, who also was involved in the study design. Logistic regression, random forest, extreme gradient boosting, and multilayer perceptron were developed and their performance of predicting improvement and deterioration in HRQOL was evaluated with accuracy, recall, specificity, and area under the ROC curve (AUC). Model performance was generally better when predicting improvement, with best models giving AUC of 0.904 for predicting mobility improvement at 12 weeks and AUC of 0.898 for predicting usual activities improvement at 18 weeks. The results encourage involving stakeholders in research and support the view that ML can be used to predict outcomes meaningful to patients. They also highlight that although some outcome variables can be valuable for patients, they may not be predicted well by ML models. This study can inform future work on patient-centric ML methods contributing to treatment decisions in oncology.

Keywords: Machine Learning · Patient-Centric Approach · Patient-Reported Data · Cancer Outcomes · Health-Related Quality Of Life

1 Introduction

Cancer diagnosis can be devastating for an individual and lead to emotional distress, anxiety or depression. Patients are usually faced with the immediate need

to make a treatment decision, which is a daunting and complicated process due to serious adverse effects of chemotherapy. Often the treatment choice requires compromising between the quality of life and the length of life [1]. There are many existing studies on cancer survival analysis, but health-related quality of life (HRQOL) changes require more consideration in literature. Understanding factors related with increased chemotherapy toxicity and HRQOL fluctuations could personalise cancer care through enabling informed shared decision-making process and early preparation for potential adverse effects [2].

Machine learning (ML) has been successful in predicting patient outcomes in healthcare [3], but ML models are often built on clinical and demographic data, overlooking patients' views. Patient-reported outcome measures (PROMs) are questionnaires collecting patients' perception on their own health status, unaffected by clinical opinion [4]. PROMs added as input data in ML models have a potential to improve their predictive performance [5]. Furthermore, patients' perspective is often missed from the study design process. Therefore, the development of new ML models might not serve its purpose or address needs of diverse groups of people in an equal and fair way. Patient-centric research is also crucial for building public trust in artificial intelligence and support the implementation of the studied tools in clinical practice [6].

There is existing evidence that ML can predict chemotherapy toxicity, but research papers often focus on acute hospital utilisation during treatment, rather than HRQOL [7]. Studies foreseeing HRQOL during chemotherapy tend to use statistical modelling, rather than explore ML predictive opportunities [8]. Furthermore, the outcomes are usually limited to specific time during the chemotherapy, without investigating the changes from early to late stages of the treatment [9]. Patient and clinician involvement is also missing from the study design.

This study is a part of a broader research on using PROMs and ML to predict patient outcomes. In earlier work we have shown that baseline PROMs can predict hospital utilisation and treatment management [7]. In this paper we *focus on predicting patients' quality of life changes during chemotherapy following a patient-centric study design*. The study utilises PROMs in ML models to predict changes in physical well-being, social functioning, role functioning, usual activities, and mobility at 6, 12 and 18 weeks of chemotherapy. The paper ensures rigorous reporting of model development and evaluation, supporting reproducibility of research [10]. Stakeholder engagement was an essential part of the study design and model evaluation.

2 Methods

2.1 Overall Methodology

Four machine learning models: logistic regression (LR), random forest (RF), extreme gradient boosting (XGB), and multilayer perceptron (MLP) were used to predict patient-reported HRQOL changes at 6, 12 and 18 weeks of starting chemotherapy treatment. The overall study design is presented in Fig. 1.

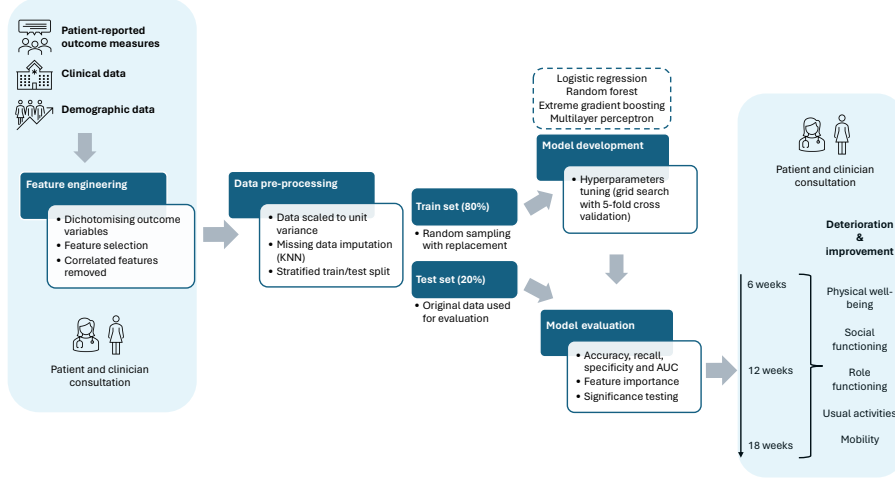


Fig. 1. Flow diagram illustrating the overall methodology of the study.

2.2 Patient-Centric Approach

A consultation with patient representatives was a part of a Data-driven Cancer Research Conference 2024 in Manchester, UK, where initial study design was presented. Two of the patient representatives were asked which chemotherapy effects would be the most helpful to be aware of before treatment decisions. Patients stressed that apart from physical symptoms, they would like to understand how chemotherapy would impact their every-day activities and social life. Consequently, these aspects were selected as outcome variables in this study.

The patients' perspective was further considered, as a clinical oncologist was involved in all stages of the study. Patient Centred Outcomes Research (PCOR) Group at the Leeds Institute of Medical Research, University of Leeds, St James's University Hospital, a multidisciplinary research group, which includes oncologists, nurses and psychologists, was also consulted during earlier project stages. Following discussions with the group, the study includes feature importance analysis and ML model evaluation on original (unprocessed) data.

2.3 Dataset

The dataset contains data from 508 patients initiating chemotherapy for colorectal, breast, or gynecological cancers at Leeds Cancer Centre, collected in a prospective, randomized two-arm parallel group study, called eRAPID clinical trial [11]. It consists of 90 variables, 42 of which are demographic (self-reported) and clinical (from Electronic Healthcare Records). They include age at study entry, sex, marital status, level of education, employment status, body mass index (BMI), disease site (breast/gynecological/colorectal), previous chemother-

apy (yes/no), disease type (metastatic/non-metastatic), and comorbidities presence: cardiovascular, respiratory, gastrointestinal, stomach/intestine, endocrine, renal, neurological, rheumatologic, previous malignancy, substance abuse. The remaining 48 variables consist of PROMs, completed by participants at 4 time-points (baseline at the start of the trial, and at 6-, 12-, and 18-week follow-up). For each time-point, 6 variables were from Five-dimensional Visual Analogue Scale (EQ-5D-VAS [12]), including self-reported data on mobility, self-care, usual activities, pain/discomfort, anxiety/depression, and self-rated health status; 4 variables from Functional Assessment of Cancer Therapy - General 28 items (FACT-G [13]), including aggregated scores of physical, social, emotional and functional well-being; and 2 remaining variables were social and role scale from EORTC Core Quality of Life Questionnaire (QLQ-C30 [14]).

2.4 Feature Engineering

The outcome variables were improvement and deterioration in: physical well-being (physical symptoms), social functioning (ability to engage in the society), role functioning (ability to perform life roles), usual activities (ability to perform daily life activities), and mobility (ability to move) at 6, 12 and 18 weeks. Baseline values were subtracted from the values at the predicted time point. The improvement and deterioration were transformed into binary variables with the threshold of minimally clinically important difference (MCID) [15,16] for physical well-being, social functioning and role functioning to ensure clinical relevance. Usual activities and mobility were 5-level, 1-item scale (from no problems to extreme problems) and have no MCID recommendations, so the deterioration and improvement of these variables were considered any change of ≤ 1 and ≤ -1 respectively (Table 1). Input features are described in *Dataset* subsection. Features correlated with Pearson coefficient higher than 0.6 were removed (leaving one) to ensure ML models process information efficiently. The list of removed variables in each feature set is presented in the Appendix Table 2.

2.5 Data Pre-Processing

Continuous variables were standardised. Only rows (participants) with complete target variable at the predicted time point were included in the analysis. Any missing data were imputed using KNN imputer ($k=5$) [17]. The number of missing data for each variable in each outcome is presented in Appendix Table 3. The data was split with to stratification train (80%) and test (20%). Random sampling with replacement was performed on training set to ensure that models are not biased towards one class [18]. Test set was left imbalanced and models were evaluated on unprocessed data to have a potential to be applied to clinical practice. Original class distribution is provided in Table 1.

2.6 Model Development and Evaluation

LR, RF, XGB and MLP were applied using Python sklearn library to predict each target variable at each time point. The model selection was based on their

Table 1. Predicted variables, questionnaires completed, change calculations, participants number (N), including deterioration (D) and improvement (I) cases.

HRQOL	Questionnaire	Deterioration	Improvement	Time point (N, D, I)
Physical well-being	Fact-G	change ≤ -2	change ≥ 2	6 weeks (N=439, D=252, I=68)
				12 weeks (N=400, D=241, I=51)
				18 weeks (N=382, D=219, I=58)
Social functioning	QLQ-C30	change ≤ -7	change ≥ 8	6 weeks (N=440, D=187, I=101)
				12 weeks (N=407, D=209, I=73)
				18 weeks (N=388, D=179, I=90)
Role functioning	QLQ-C30	change ≤ -6	change $\geq 11^*$	6 weeks (N=438, D=190, I=100)
				12 weeks (N=407, D=216, I=90)
				18 weeks (N=385, D=171, I=78)
Usual activities	EQ-5D-VAS	change ≥ 1	change ≥ -1	6 weeks (N=439, D=148, I=71)
				12 weeks (N=408, D=181, I=50)
				18 weeks (N=384, D=183, I=51)
Mobility	EQ-5D-VAS	change ≥ 1	change ≥ -1	6 weeks (N=438, D=64, I=48)
				12 weeks (N=408, D=98, I=40)
				18 weeks (N=382, D=122, I=31)

*MCID for role functioning improvement was the average from all cancer types due to availability [16].

frequency of use in studies applying ML to PROMs data to enable between-studies comparison. Hyperparameters were tuned on training sets through grid search with five-fold cross-validation. The models were evaluated with accuracy, recall (also known as sensitivity), specificity and area under the ROC curve (AUC). Feature importance was also conducted on LR and RF models predicting 3 best performing outcomes at 18 weeks due to their explainability potential. Furthermore, Analysis of Variance (ANOVA) with Tukey’s Honest Significant Difference (Tukey HSD) tests were performed to compare model performances.

3 Results

Accuracy, recall, specificity and AUC of all models are provided in Appendix Tables 4 and 5. The plots in Fig. 2 represent AUCs of all target variables apart from role functioning, due to its similarity to usual activities, but worse performance.

3.1 Predicted HRQOL Changes

Physical Well-Being. *Deterioration:* LR models had the highest AUCs at 6 (0.710), 12 (0.661) and 18 (0.761) weeks deterioration predictions. Even though LR did not provide the highest recall, it did not compromise the specificity unlike other models (e.g., MLP at 6 weeks: recall=0.902, specificity=0.162), preventing false negative predictions. All models had the best performance predicting deteriorations at 18 weeks, compared to 6 and 12 weeks. *Improvement:* MLP had the highest AUC overall at 12 weeks (0.843). This model also had high recall (0.800) and specificity (0.886). LR had the highest performance at 6 (AUC=0.718) and 18 weeks (AUC=0.662). Generally all models had poor recall (risking false improvement predictions), apart from LR (0.800) and aforementioned MLP (0.800) predicting improvement at 12 weeks. The models performed the best at 12 weeks, except for RF achieving highest AUC at 6 weeks.

Social Functioning. *Deterioration:* Overall, the best performing model was LR with highest AUC of 0.685 at 18 weeks. MLP was the best classifier at 6 weeks (AUC=0.595), and LR at 12 weeks (AUC=0.562). The models were compromising either specificity or recall, with at least one of them being lower than 0.600. Overall, 18 weeks deterioration predictions had highest AUC for all models except MLP (highest AUC of 0.595 at 6 weeks). *Improvement:* The best performing model was MLP at 6 (AUC=0.713), 12 (AUC=0.533) and 18 (AUC=0.736) weeks. These models also had highest recall, which was 0.750 at 6 weeks and 0.722 at 18 weeks with specificity being respectively 0.676 and 0.750. The MLP’s recall at 12 weeks was only 0.200, suggesting a bad performance of this model. RF had the highest specificity for all time points, but compromised recall. All models achieved best predictive performance at 18 weeks.

Role Functioning. *Deterioration:* LR outperformed other models, achieving the highest AUCs at all time points: 6 (0.573), 12 (0.691) and 18 (0.675) weeks. Recalls of these models were higher than specificity, with the highest value at 12 weeks (0.750). The models had best overall performance at 18 weeks, even though LR was slightly better at 12 weeks. *Improvement:* LR had the highest AUCs of all models at 6 (0.531) and 18 weeks (0.770), whilst MLP at 12 weeks (0.712). LR’s performance at 18 weeks was the highest overall. This model also had the highest recall (0.688), but specificity of 0.852 was lower than XGB’s specificity (0.902). Overall, models performed the best at 18 weeks (except MLP with the highest AUC=0.712 at 12 weeks) and the worst at 6 weeks.

Usual Activities. *Deterioration:* The best performing model was LR at 6 (AUC=0.663) and 12 weeks (AUC=0.739), and MLP (AUC=0.718) at 18 weeks. These models also had highest recall, but not specificity, which was the highest for RF (6 and 12 weeks) and XGB (18 weeks). In general, specificity was a lot higher than recall. The models had the lowest performance at 6 weeks, while 12 and 18 weeks predictions had similar performance. *Improvement:* LR outperformed all models at all time points with AUCs of 0.684 (6 weeks), 0.731 (12 weeks), and 0.898 (18 weeks). LR at 18 weeks was the highest performing model with excellent recall (0.900) and specificity (0.896). The specificity of all models was high, but often compromised recall. Models predicting 18 weeks improvement performed much better than models at other time points.

Mobility. *Deterioration:* The models had poor overall performance with MLP having the highest AUCs at 6 (0.529) and 12 (0.562) weeks, and LR (0.597) at 18 weeks. Recall of the models was poor as well, with the highest value of 0.560. Specificity values of the models were good overall, with the highest for RF at 6 weeks (0.973). It is unclear which time point resulted in the best predictions. *Improvement:* MLP was the best performing model overall achieving AUC of 0.904 (recall=0.875, specificity=0.932) at 12 weeks. LR was the best models at 6 weeks (AUC=0.829) with very good recall (0.800) and specificity (0.859). At 18

weeks LR also outperformed other models with AUC of 0.715, high specificity (0.930), but low recall (0.500). Overall, 18 weeks predictions resulted in the lowest AUCs, while 6- and 12-week predictions depended on the model.

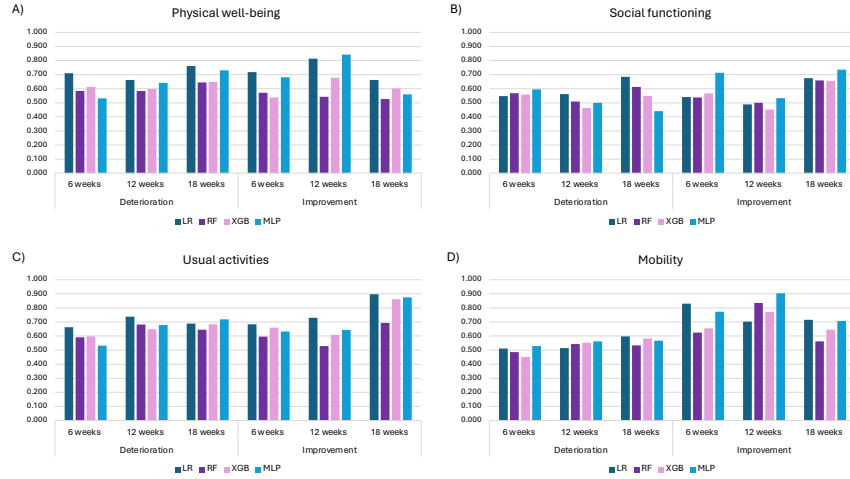


Fig. 2. AUC values of models predicting A) physical well-being, B) social functioning, C) usual activities, D) mobility deterioration and improvement at 6, 12, and 18 weeks.

3.2 Significance Testing

ANOVA results suggest that at least one model's AUC mean is different from the rest ($F=4.202$, $p=0.007$) and Tukey HSD indicated that LR was significantly better than RF (0.011) and XGB (0.041). Outcome also affected prediction performance ($F=4.659$, $p=0.002$), as predictions of usual activities had higher mean AUC than role ($p=0.0363697$) and social ($p=0.001$) functioning. The time point of prediction was also a factor affecting AUC ($F=5.676$, $p=0.004$), with Tukey HSD indicating 18-weeks predictions resulting with significantly higher mean AUC than 6-weeks predictions ($p=0.003$). Welch Two Sample t-test indicated that improvement predictions were significantly better than deterioration predictions ($t=-3.079$, $df=101.39$, $p=0.003$).

3.3 Feature Importance

Features with coefficients for LR and values of importance for RF are presented in Appendix Tables 6,7,8. For prediction of physical well-being, LR looked mainly at clinical information (comorbidities and cancer characteristics). RF considered Fact-G and EQ-5D as most important predictors, as well as patients' BMI. LR

in predicting mobility also found comorbidities the most meaningful predictors (including EQ5D mobility score at baseline for improvement prediction), while RF mainly looked at PROMs, cancer type and BMI. When predicting usual activities, LR also looked at the clinical characteristics and usual activities baseline score, whilst RF focused on BMI, cancer type and PROMs.

4 Discussion

4.1 Discussion of the Study Findings

Overall results. The findings suggest that ML models applied on PROMs, clinical and demographic data can successfully predict HRQOL outcomes throughout chemotherapy which are meaningful to patients. Models provided excellent performance in predicting improvement in physical well-being, usual activities and mobility at different stages of cancer treatment. Consultation with a clinical oncologist endorsed the view that ML prediction of HRQOL changes during chemotherapy can be useful in clinical practice.

Model performances. LR generally outperformed other models, which is a common outcome in medical research, as other models are more susceptible to overfitting [19]. In some cases, MLP achieved higher performance than LR. Nevertheless, the lack of explainability of MLP could affect public trust in this model. According to the consulted clinical oncologist, even impossible to interpret models should be considered as useful, as long as they are used alongside other well-performing models, which enable explainability.

Change at given time points. Improvement predictions had generally higher performance than deterioration, even though chemotherapy is associated with decline in HRQOL [8]. However, the deterioration might depend on patient characteristics from the start of chemotherapy. For example, metastatic disease may be more likely to show improvement due to higher burden of cancer symptoms prior to treatment. These differences will be further explored.

4.2 Strengths and Limitations

The main strength of this work is the patient-centric approach achieved through the active engagement of a clinical oncologist, consultation with patient representatives and patient reports used as input data. Three time points of HRQOL changes provided another insight into fluctuation of chemotherapy symptoms and when they can affect individuals. Finally, the rigorous reporting of data pre-processing methods, model development and evaluation supports the reproducibility of this study. However, the ethical approval does not allow data sharing, which might negatively impact the reproducibility. Furthermore, this study has limitations typical for data collected in clinical trials. While the ML models have been rigorously designed and evaluated, the data were subject to inclusion and exclusion criteria, which can lead to bias. Temporal clinical trial data are usually relatively small samples and are affected by participants drop outs. This limits the ML methods that have been applied, yet the results are encouraging.

4.3 Conclusions and Future Work

This study successfully applied ML models on PROMs, clinical and demographic data to predict changes in HRQOL during chemotherapy, which could support preparation for adverse effects of chemotherapy and inform treatment decisions. The results further encourage the use of ML methods to identify factors related to chemotherapy toxicity and explore how cancer treatment affects individuals' lives. Patient and clinician involvement ensured that the predicted variables are meaningful for patients and clinically relevant. We are currently extending the stakeholder engagement by designing ways to explain ML models and evaluate possible clinical adoption of the findings from a more representative group of patients. We are also using longitudinally collected PROMs and symptom reports for patient outcome predictions. This will consider traditional ML models and deep learning methods to process multi-dimensional time-series data.

Acknowledgements. The authors thank the patients and clinicians participating in the eRAPID clinical trial for providing the data; Patient Centred Outcomes Research (PCOR) Group in the University of Leeds, Faculty of Medicine and Health, Leeds Institute of Medical Research for consultations of initial study design; and patient representatives of the Use My Data for discussing research relevance. This work was supported in part by UK Research and Innovation (UKRI) [CDT grant number EP/S024336/1].

References

1. M. Kowal, F. Douglas, D. Jayne, and D. Meads, "Patient choice in colorectal cancer treatment – A systematic review and narrative synthesis of attribute-based stated preference studies," *Colorectal Disease*, vol. 24, no. 11, pp. 1295–1307, 2022.
_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/codi.16242>.
2. W. Xuyi, H. Seow, and R. Sutradhar, "Artificial neural networks for simultaneously predicting the risk of multiple co-occurring symptoms among patients with cancer," *Cancer Medicine*, vol. 10, no. 3, pp. 989–998, 2021.
_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cam4.3685>.
3. M. Shehab, L. Abualigah, Q. Shambour, M. A. Abu-Hashem, M. K. Y. Shambour, A. I. Alsalibi, and A. H. Gandomi, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, June 2022.
4. C. Kingsley and S. Patel, "Patient-reported outcome measures and patient-reported experience measures," *BJA Education*, vol. 17, pp. 137–144, Apr. 2017.
5. J.-a. Sim, Y. A. Kim, J. H. Kim, J. M. Lee, M. S. Kim, Y. M. Shim, J. I. Zo, and Y. H. Yun, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Scientific Reports*, vol. 10, p. 10693, July 2020.
6. Y. Chen, A. A. Hosin, M. J. George, F. W. Asselbergs, and A. D. Shah, "Digital technology and patient and public involvement (PPI) in routine care and clinical research—A pilot study," *PLOS ONE*, vol. 18, p. e0278260, Feb. 2023.

7. Z. Wójcik, Vania Dimitrova, Lorraine Warrington, Galina Velikova, and Kate Absolom, "Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From Patient- Reported Outcome Measures," *JCO Clinical Cancer Informatics (In press)*, 2024.
8. K. Zhou, M. Bellanger, S. Le Lann, M. Robert, J.-S. Frenel, and M. Campone, "The predictive value of patient-reported outcomes on the impact of breast cancer treatment-related quality of life," *Frontiers in Oncology*, vol. 12, p. 925534, 2022.
9. Y. Wang, L. Van Dijk, A. S. R. Mohamed, C. D. Fuller, X. Zhang, G. E. Marai, and G. Canahuate, "Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes," *Proceedings. International Database Engineering and Applications Symposium*, vol. 2021, pp. 273–279, July 2021.
10. D. Jha, A. Rauniyar, A. Srivastava, D. H. Hagos, N. K. Tomar, V. Sharma, E. Kelles, Z. Zhang, U. Demir, A. Topcu, A. Yazidi, J. E. Håakegård, and U. Bagci, "Ensuring Trustworthy Medical Artificial Intelligence through Ethical and Philosophical Principles," Sept. 2023. arXiv:2304.11530 [cs].
11. K. Absolom, L. Warrington, E. Hudson, J. Hewison, C. Morris, P. Holch, R. Carter, A. Gibson, M. Holmes, B. Clayton, Z. Rogers, L. McParland, M. Conner, L. Glidewell, B. Woroncow, B. Dawkins, S. Dickinson, C. Hulme, J. Brown, and G. Velikova, "Phase III Randomized Controlled Trial of eRAPID: eHealth Intervention During Chemotherapy," *Journal of Clinical Oncology*, vol. 39, pp. 734–747, Mar. 2021. Publisher: Wolters Kluwer.
12. P. Dolan, "Modeling valuations for EuroQol health states," *Medical Care*, vol. 35, pp. 1095–1108, Nov. 1997.
13. D. F. Cella, D. S. Tulsky, G. Gray, B. Sarafian, E. Linn, A. Bonomi, M. Silberman, S. B. Yellen, P. Winicour, and J. Brannon, "The Functional Assessment of Cancer Therapy scale: development and validation of the general measure.," *Journal of Clinical Oncology*, vol. 11, pp. 570–579, Mar. 1993. Publisher: Wolters Kluwer.
14. N. K. Aaronson, S. Ahmedzai, B. Bergman, M. Bullinger, A. Cull, N. J. Duez, A. Filiberti, H. Flechtner, S. B. Fleishman, and J. C. de Haes, "The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology," *Journal of the National Cancer Institute*, vol. 85, pp. 365–376, Mar. 1993.
15. M. T. King, M. R. Stockler, D. F. Cella, D. Osoba, D. T. Eton, J. Thompson, and A. R. Eisenstein, "Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G," *Journal of Clinical Epidemiology*, vol. 63, pp. 270–281, Mar. 2010.
16. J. Z. Musoro, C. Coens, F. Fiteni, P. Katarzyna, F. Cardoso, N. S. Russell, M. T. King, K. Cocks, M. A. Sprangers, M. Groenvold, G. Velikova, H.-H. Flechtner, and A. Bottomley, "Minimally Important Differences for Interpreting EORTC QLQ-C30 Scores in Patients With Advanced Breast Cancer," *JNCI Cancer Spectrum*, vol. 3, p. pkz037, June 2019.
17. T. Aljrees, "Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning," *PLOS ONE*, vol. 19, p. e0295632, Jan. 2024.
18. R. Shafique, F. Rustam, G. S. Choi, I. d. l. T. Díez, A. Mahmood, V. Lipari, C. L. R. Velasco, and I. Ashraf, "Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning," *Cancers*, vol. 15, p. 681, Jan. 2023.
19. E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, June 2019.

A Appendix

Table 2. Removed features based on correlation analysis.

Physical well-being	6 weeks	'SOC0', 'ROL0', 'AgeStudyEntry', 'Sex', 'EWB_overall0'
	12 weeks	'FWB_overall0', 'SOC0', 'ROL0', 'AgeStudyEntry', 'Sex', 'EWB_overall0', 'CMRheuCTD'
	18 weeks	'AgeStudyEntry', 'EWB_overall0', 'Sex', 'SOC0'
Social functioning	6 weeks	'SOC0', 'EWB_overall0', 'AgeStudyEntry', 'ROL0', 'Sex'
	12 weeks	'SOC0', 'AgeStudyEntry', 'EWB_overall0', 'CMRheuCTD', 'CMCarHyperten', 'FWB_overall0', 'ROL0', 'Sex'
	18 weeks	'SOC0', 'EWB_overall0', 'AgeStudyEntry', 'Sex'
Role functioning	6 weeks	'SOC0', 'ROL0', 'Sex', 'AgeStudyEntry', 'FWB_overall0', 'EWB_overall0'
	12 weeks	'SOC0', 'ROL0', 'Sex', 'CMRheuCTD', 'AgeStudyEntry', 'FWB_overall0', 'EWB_overall0', 'CMCarHyperten'
	18 weeks	'SOC0', 'Sex', 'AgeStudyEntry', 'EWB_overall0', 'CMCarHyperten'
Usual activities	6 weeks	'ROL0', 'EWB_overall0', 'FWB_overall0', 'Sex', 'AgeStudyEntry', 'SOC0'
	12 weeks	'ROL0', 'EWB_overall0', 'CMCarHyperten', 'FWB_overall0', 'CMRheuCTD', 'Sex', 'AgeStudyEntry', 'SOC0'
	18 weeks	'EWB_overall0', 'FWB_overall0', 'Sex', 'AgeStudyEntry', 'SOC0'
Mobility	6 weeks	'AgeStudyEntry', 'FWB_overall0', 'SOC0', 'ROL0', 'Sex', 'FWB_overall0', 'EWB_overall0'
	12 weeks	'AgeStudyEntry', 'ROL0', 'SOC0', 'Sex', 'CMCarHyperten', 'FWB_overall0', 'CMRheuCTD', 'EWB_overall0'
	18 weeks	'AgeStudyEntry', 'SOC0', 'ROL0', 'Sex', 'EWB_overall0'

Table 3. Number of missing data in each variable for each outcome at 18 weeks.

Physical well-being		Social functioning		Role functioning		Usual activities		Mobility	
StudyArm	0	StudyArm	0	StudyArm	0	StudyArm	0	StudyArm	0
DiseaseSite	0	DiseaseSite	0	DiseaseSite	0	DiseaseSite	0	DiseaseSite	0
Sex	0	Sex	0	Sex	0	Sex	0	Sex	0
PreviousChemo	0	PreviousChemo	0	PreviousChemo	0	PreviousChemo	0	PreviousChemo	0
AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0	AgeStudyEntry	0
PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0	PrimaryorMet	0
BCBMI	1	BCBMI	1	BCBMI	1	BCBMI	1	BCBMI	2
Comorbidities	0	Comorbidities	0	Comorbidities	0	Comorbidities	0	Comorbidities	0
CMCarMI	0	CMCarMI	0	CMCarMI	0	CMCarMI	0	CMCarMI	0
CMCarAngina	0	CMCarAngina	0	CMCarAngina	0	CMCarAngina	0	CMCarAngina	0
CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0	CMCarHeartFail	0
CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0	CMCarArrhythm	0
CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0	CMCarHyperten	0
CMCarVenous	0	CMCarVenous	0	CMCarVenous	0	CMCarVenous	0	CMCarVenous	0
CMResCOPD	0	CMResCOPD	0	CMResCOPD	0	CMResCOPD	0	CMResCOPD	0
CMResEmphys	0	CMResEmphys	0	CMResEmphys	0	CMResEmphys	0	CMResEmphys	0
CMResAsthma	0	CMResAsthma	0	CMResAsthma	0	CMResAsthma	0	CMResAsthma	0
CMResChronBron	0	CMResChronBron	0	CMResChronBron	0	CMResChronBron	0	CMResChronBron	0
CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0	CMGasChronHep	0
CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0	CMGasCirrhosis	0
CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0	CMGasPancreas	0
CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0	CMStomUlcers	0
CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0	CMStomMalabsor	0
CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0	CMStomInflamm	0
CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0	CMEndDiabetes	0
CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0	CMEndHypothy	0
CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0	CMEndHyperth	0
CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0	CMRenEndStage	0
CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0	CMNeuStroke	0
CMNeuMS	0	CMNeuMS	0	CMNeuMS	0	CMNeuMS	0	CMNeuMS	0
CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0	CMNeuParkins	0
CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0	CMNeuMyasth	0
CMRheuArth	0	CMRheuArth	0	CMRheuArth	0	CMRheuArth	0	CMRheuArth	0
CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0	CMRheuLupus	0
CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0	CMRheuCTD	0
CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0	CMRheuPolymyo	0
CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMRheuRhPolymy	0
CMPrevMal	0	CMPrevMal	0	CMPrevMal	0	CMPrevMal	0	CMPrevMal	0
CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0	CMSubstAlcohol	0
CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0
DCMarital	4	DCMarital	4	DCMarital	4	DCMarital	4	DCMarital	4
DCEmployment	12	DCEmployment	12	DCEmployment	12	DCEmployment	12	DCEmployment	12
ed_lev	12	ed_lev	12	ed_lev	12	ed_lev	12	ed_lev	12
EQ5DMob0	4	EQ5DMob0	4	EQ5DMob0	4	EQ5DMob0	3	EQ5DMob0	2
EQ5DSelCar0	3	EQ5DSelCar0	3	EQ5DSelCar0	3	EQ5DSelCar0	2	EQ5DMob18	0
EQ5DUsuAct0	3	EQ5DUsuAct0	3	EQ5DUsuAct0	3	EQ5DUsuAct0	2	EQ5DSelCar0	2
EQ5DPain0	5	EQ5DPain0	5	EQ5DPain0	5	EQ5DUsuAct18	0	EQ5DUsuAct0	2
EQ5DAnxDep0	5	EQ5DAnxDep0	5	EQ5DAnxDep0	5	EQ5DPain0	4	EQ5DPain0	4
EQ5DVAS0	4	EQ5DVAS0	4	EQ5DVAS0	4	EQ5DAnxDep0	4	EQ5DAnxDep0	4
ROL0	2	ROL0	2	ROL0	2	EQ5DVAS0	3	EQ5DVAS0	3
SOC0	2	SOC0	2	ROL18	0	ROL0	2	ROL0	3
PWB_overall0	8	SOC18	0	SOC0	2	SOC0	2	SOC0	3
PWB_overall18	0	PWB_overall0	8	PWB_overall0	8	PWB_overall0	8	PWB_overall0	9
SWB_overall0	4	SWB_overall0	4	SWB_overall0	4	SWB_overall0	4	SWB_overall0	5
EWB_overall0	1	EWB_overall0	1	EWB_overall0	1	EWB_overall0	1	EWB_overall0	2
FWB_overall0	1	FWB_overall0	1	FWB_overall0	1	FWB_overall0	1	FWB_overall0	2

Table 4. Deterioration prediction results with hyperparameters used for model development.

Data		Deterioration					
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters
Physical well-being	6 weeks	LR	0.716	0.676	0.745	0.710	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.625	0.324	0.843	0.584	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	0.636	0.459	0.765	0.612	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
	12 weeks	MLP	0.591	0.162	0.902	0.532	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
		LR	0.663	0.656	0.667	0.661	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.613	0.438	0.729	0.583	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
	18 weeks	XGB	0.625	0.469	0.729	0.599	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 100
		MLP	0.638	0.656	0.625	0.641	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
		LR	0.766	0.727	0.795	0.761	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.675	0.421	0.864	0.644	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	0.662	0.545	0.750	0.648	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	0.740	0.667	0.795	0.751	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters
Social functioning	6 weeks	LR	0.557	0.698	0.486	0.547	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.602	0.784	0.351	0.568	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	0.591	0.765	0.351	0.558	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
	12 weeks	MLP	0.591	0.569	0.622	0.598	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
		LR	0.561	0.600	0.524	0.562	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.512	0.375	0.643	0.509	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
	18 weeks	XGB	0.463	0.425	0.500	0.463	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	0.512	0.000	1.000	0.500	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
		LR	0.692	0.786	0.583	0.685	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.628	0.810	0.417	0.613	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	0.564	0.738	0.381	0.550	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	0.436	0.381	0.500	0.440	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'sgd'
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters
Role functioning	6 weeks	LR	0.568	0.540	0.605	0.573	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	0.580	0.820	0.263	0.542	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	0.591	0.780	0.342	0.561	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
	12 weeks	MLP	0.432	0.000	1.000	0.500	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'sgd'
		LR	0.695	0.632	0.759	0.691	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.646	0.646	0.841	0.631	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
	18 weeks	XGB	0.573	0.395	0.727	0.561	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	0.524	0.395	0.636	0.516	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
		LR	0.675	0.674	0.676	0.676	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	0.675	0.837	0.471	0.654	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	0.636	0.721	0.529	0.625	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 100
		MLP	0.662	0.628	0.706	0.667	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters
Usual activities	6 weeks	LR	0.693	0.759	0.567	0.663	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	0.705	0.948	0.233	0.591	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	0.670	0.828	0.367	0.597	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
	12 weeks	MLP	0.659	0.931	0.133	0.532	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
		LR	0.744	0.783	0.694	0.739	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.707	0.801	0.472	0.682	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
	18 weeks	XGB	0.671	0.826	0.472	0.649	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	0.683	0.717	0.639	0.678	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
		LR	0.688	0.675	0.703	0.689	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	0.649	0.750	0.541	0.645	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	0.688	0.800	0.568	0.684	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	0.714	0.625	0.811	0.718	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters
Mobility	6 weeks	LR	0.545	0.560	0.462	0.511	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	0.830	0.973	0.000	0.487	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	0.716	0.827	0.077	0.432	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
	12 weeks	MLP	0.739	0.827	0.231	0.526	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
		LR	0.549	0.581	0.450	0.515	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	0.744	0.935	0.150	0.543	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
	18 weeks	XGB	0.683	0.860	0.300	0.555	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	0.671	0.774	0.350	0.562	activation: 'relu', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
		LR	0.610	0.635	0.560	0.597	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'ldgs'
		RF	0.625	0.788	0.280	0.534	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	0.662	0.808	0.360	0.584	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	0.584	0.615	0.520	0.568	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'

Table 5. Improvement prediction results with hyperparameters used for model development.

Data		Improvement					Hyperparameters	
Outcome	Outcome at	Model	Accuracy	Specificity	Recall	AUC		
Physical well-being	6 weeks	LR	0.818	0.865	0.571	0.718	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'	
		RF	0.864	1.000	0.143	0.571	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500	
		XGB	0.807	0.932	0.143	0.538	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 100	
		MLP	0.852	0.932	0.429	0.681	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.825	0.829	0.800	0.814	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'lbfgs'	
		RF	0.875	0.986	0.100	0.543	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
	12 weeks	XGB	0.888	0.957	0.400	0.799	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.875	0.886	0.800	0.843	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.831	0.908	0.417	0.662	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'lbfgs'	
		RF	0.831	0.960	0.083	0.526	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.844	0.954	0.250	0.602	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500	
		MLP	0.831	0.954	0.167	0.560	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
	18 weeks	Model	Accuracy	Specificity	Recall	AUC	Hyperparameters	
		LR	0.591	0.632	0.450	0.541	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'	
		RF	0.750	0.926	0.150	0.538	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.739	0.882	0.250	0.566	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500	
		MLP	0.693	0.676	0.750	0.713	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.671	0.776	0.200	0.488	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
Social functioning	6 weeks	RF	0.817	1.000	0.000	0.500	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.695	0.695	0.067	0.451	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.714	0.896	0.200	0.533	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.679	0.683	0.667	0.675	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.833	0.983	0.333	0.658	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500	
		XGB	0.769	0.867	0.444	0.656	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500	
	12 weeks	MLP	0.744	0.750	0.722	0.736	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
		Model	Accuracy	Specificity	Recall	AUC	Hyperparameters	
		LR	0.602	0.662	0.400	0.531	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'	
		RF	0.750	0.941	0.100	0.521	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.659	0.824	0.100	0.462	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.523	0.574	0.350	0.462	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
	18 weeks	LR	0.720	0.734	0.667	0.701	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.756	0.938	0.111	0.524	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.720	0.875	0.167	0.521	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500	
		MLP	0.768	0.813	0.611	0.712	activation: 'tanh', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.818	0.852	0.688	0.770	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.766	0.869	0.375	0.622	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
Role functioning	6 weeks	XGB	0.792	0.902	0.375	0.638	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.714	0.721	0.688	0.704	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		Model	Accuracy	Specificity	Recall	AUC	Hyperparameters	
		LR	0.761	0.797	0.571	0.684	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'	
		RF	0.807	0.905	0.286	0.596	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.818	0.897	0.429	0.660	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500	
	12 weeks	MLP	0.773	0.838	0.429	0.633	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.829	0.861	0.600	0.731	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.854	0.958	0.100	0.529	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.841	0.917	0.300	0.608	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.829	0.889	0.400	0.644	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.896	0.896	0.900	0.898	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'	
	18 weeks	RF	0.909	0.985	0.400	0.693	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.909	0.925	0.800	0.863	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 100	
		MLP	0.857	0.851	0.900	0.875	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
		Model	Accuracy	Specificity	Recall	AUC	Hyperparameters	
		LR	0.852	0.859	0.800	0.820	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.875	0.949	0.300	0.624	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100	
Mobility	6 weeks	XGB	0.852	0.910	0.400	0.655	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.831	0.846	0.700	0.773	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'	
		LR	0.866	0.905	0.500	0.703	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.902	0.919	0.750	0.834	criterion: 'gini', 'max_depth': 8, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.890	0.910	0.625	0.774	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 100	
		MLP	0.927	0.932	0.875	0.904	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	
	12 weeks	LR	0.896	0.930	0.500	0.715	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'bilinear'	
		RF	0.896	0.958	0.167	0.562	criterion: 'gini', 'max_depth': 8, 'min_samples_split': 2, 'n_estimators': 100	
		XGB	0.909	0.958	0.333	0.646	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500	
		MLP	0.883	0.915	0.500	0.708	activation: 'relu', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'	

Table 6. Feature importance ranks for physical well-being changes at 18 weeks prediction models LR and RF.

Physical well-being							
Deterioration				Improvement			
Logistic regression		Random forest		Logistic regression		Random forest	
PWB_overall0	1.569433	PWB_overall0	0.133532	CMResCOPD	13.923359	PWB_overall0	0.221759
CMRheuArth	1.149526	BCBMI	0.09927	CMCarMI	13.066001	BCBMI	0.073412
CMStomInflamm	1.026601	FWB_overall0	0.081032	CMCarAngina	9.672283	EQ5DPain0	0.071416
PrimaryorMet	0.879178	SWB_overall0	0.075772	CMEndDiabetes	4.327627	EQ5DVAS0	0.064324
PreviousChemo	0.864974	EQ5DVAS0	0.066705	CMEndHyperth	4.285474	SWB_overall0	0.062971
CMPrevMal	0.79102	DiseaseSite	0.050577	CMNeuStroke	3.695441	FWB_overall0	0.062633
CMResCOPD	0.764014	DCEmployment	0.050358	CMEndHypoathy	3.601606	ROL0	0.050095
CMCarMI	0.59501	ROL0	0.050047	Comorbidities	2.922004	DiseaseSite	0.045489
DiseaseSite	0.564434	PrimaryorMet	0.041654	CMCarHyperten	2.919881	ed_lev	0.038387
CMEndDiabetes	0.501928	ed_lev	0.039405	PWB_overall0	2.850354	PrimaryorMet	0.035688
CMRheuCTD	0.444377	EQ5DAnxDep0	0.034651	CMCarVenous	2.747976	DCEmployment	0.03415
FWB_overall0	0.444371	DCMarital	0.030691	CMCarArrhythm	2.534306	DCMarital	0.028505
CMEndHypoathy	0.426059	EQ5DPain0	0.028751	PreviousChemo	2.414958	EQ5DUsuAct0	0.026227
EQ5DAnxDep0	0.383677	EQ5DUsuAct0	0.028591	CMRheuArth	1.93644	Comorbidities	0.025013
EQ5DPain0	0.364803	Comorbidities	0.028133	PrimaryorMet	1.54155	EQ5DAnxDep0	0.024492
CMCarVenous	0.332789	EQ5DMob0	0.021961	CMResAsthma	1.499618	EQ5DMob0	0.020482
CMResAsthma	0.289044	StudyArm	0.019965	CMPrevMal	1.319404	StudyArm	0.020222
StudyArm	0.28145	PreviousChemo	0.016012	CMSubstAlcohol	1.042471	PreviousChemo	0.013142
BCBMI	0.252698	CMCarHyperten	0.015805	DiseaseSite	0.893342	CMEndDiabetes	0.013136
CMEndHyperth	0.243022	CMResCOPD	0.012084	ed_lev	0.864195	CMCarHyperten	0.01048
CMGasPancreas	0.228956	CMResAsthma	0.011007	EQ5DVAS0	0.832183	CMEndHypoathy	0.009797
ROL0	0.224995	EQ5DSelCar0	0.010529	EQ5DSelCar0	0.796896	EQ5DSelCar0	0.009597
CMStomUlcers	0.200148	CMPrevMal	0.009927	CMGasPancreas	0.74042	CMPrevMal	0.008038
CMNeuStroke	0.197854	CMEndDiabetes	0.009848	CMRheuLupus	0.703934	CMResAsthma	0.006365
CMRenEndStage	0.178679	CMCarMI	0.006526	BCBMI	0.599315	CMCarAngina	0.004794
ed_lev	0.135084	CMEndHypoathy	0.006286	FWB_overall0	0.560101	CMCarMI	0.003373
Comorbidities	0.127383	CMStomInflamm	0.004816	CMNeuParkins	0.392506	CMEndHyperth	0.002924
CMCarAngina	0.123243	CMCarVenous	0.004777	CMStomInflamm	0.349838	CMCarVenous	0.002866
EQ5DVAS0	0.110928	CMCarAngina	0.002398	DCEmployment	0.189108	CMNeuParkins	0.00223
CMCarArrhythm	0.106706	CMNeuStroke	0.001957	DCMarital	0.180942	CMResCOPD	0.001576
SWB_overall0	0.102555	CMRheuArth	0.001828	EQ5DMob0	0.171249	CMCarArrhythm	0.001128
DCEmployment	0.100668	CMRheuLupus	0.001095	CMStomUlcers	0.149692	CMRheuArth	0.001108
EQ5DMob0	0.081411	CMCarArrhythm	0.000916	ROL0	0.100497	CMStomInflamm	0.001044
CMCarHeartFail	0.078057	CMRheuCTD	0.000801	SWB_overall0	0.099762	CMNeuStroke	0.000839
CMSubstAlcohol	0.076215	CMEndHyperth	0.000681	EQ5DAnxDep0	0.085732	CMGasPancreas	0.000631
DCMarital	0.07464	CMGasPancreas	0.000615	StudyArm	0.080389	CMSubstAlcohol	0.00056
EQ5DUsuAct0	0.022877	CMSubstAlcohol	0.000385	EQ5DUsuAct0	0.042851	CMRheuCTD	0.000482
CMCarHyperten	0.012993	CMRenEndStage	0.000346	EQ5DPain0	0.042741	CMRheuLupus	0.00029
EQ5DSelCar0	0.010388	CMStomUlcers	0.000227	CMRheuCTD	0.035837	CMStomUlcers	0.000284
CMRheuLupus	0.010197	CMCarHeartFail	0.00004	CMResChronBron	0.005954	CMResChronBron	0.000052
CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMGasChronHep	0
CMResChronBron	0	CMGasChronHep	0	CMStomMalabsor	0	CMRheuRhPolymy	0
CMStomMalabsor	0	CMResChronBron	0	CMRheuRhPolymy	0	CMResEmphys	0
CMResEmphys	0	CMNeuMyasth	0	CMGasCirrhosis	0	CMRheuPolymy	0
CMRheuRhPolymy	0	CMNeuParkins	0	CMRheuPolymy	0	CMSubstDrugs	0
CMNeuMyasth	0	CMNeuMS	0	CMGasChronHep	0	CMNeuMyasth	0
CMGasChronHep	0	CMGasCirrhosis	0	CMResEmphys	0	CMNeuMS	0
CMNeuParkins	0	CMRheuRhPolymy	0	CMRenEndStage	0	CMCarHeartFail	0
CMGasCirrhosis	0	CMRheuPolymy	0	CMCarHeartFail	0	CMGasCirrhosis	0
CMNeuMS	0	CMStomMalabsor	0	CMNeuMyasth	0	CMRenEndStage	0
CMRheuPolymy	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor	0

Table 7. Feature importance ranks for usual activities changes at 18 weeks prediction models LR and RF.

Usual activities							
Deterioration			Improvement				
Logistic regression		Random forest		Logistic regression		Random forest	
EQ5DUsuAct0	2.348261	BCBMI	0.143117	CMSubstAlcohol	7.569993	EQ5DUsuAct0	0.252763
CMRheuLupus	1.244129	EQ5DUsuAct0	0.129836	CMRheuArth	3.585515	ROL0	0.115769
DiseaseSite	0.898524	DiseaseSite	0.077861	EQ5DUsuAct0	3.522745	PWB_overall0	0.08537
CMPrevMal	0.865107	EQ5DVAS0	0.07398	CMCarVenous	2.355063	BCBMI	0.08323
CMResAsthma	0.771298	SWB_overall0	0.068102	CMCarArrhythm	2.165607	EQ5DVAS0	0.063433
CMCarHyperten	0.679204	PWB_overall0	0.062277	CMStomInflamm	2.008561	SWB_overall0	0.061488
CMGasPancreas	0.663488	ROL0	0.060521	CMEndHypo	1.993451	DCEmployment	0.039528
CMEndHypo	0.542003	DCEmployment	0.045276	CMRheuLupus	1.647547	EQ5DPain0	0.032502
BCBMI	0.506767	EQ5DAnxDep0	0.0362	CMResCOPD	1.407293	DiseaseSite	0.026328
CMRenEndStage	0.498677	DCMarital	0.03589	CMRheuCTD	1.285911	ed_le	0.025657
EQ5DPain0	0.497696	ed_le	0.033316	CMCarHeartFail	1.135531	DCMarital	0.025267
EQ5DSelCar0	0.47095	EQ5DPain0	0.032078	CMCarAngina	1.051474	EQ5DAnxDep0	0.023673
CMResCOPD	0.446885	Comorbidities	0.027761	CMNeuParkins	1.035357	EQ5DMob0	0.023085
CMCarVenous	0.422538	PrimaryorMet	0.024138	PWB_overall0	0.911066	Comorbidities	0.021449
CMCarMI	0.40174	StudyArm	0.022059	CMNeuStroke	0.785979	EQ5DSelCar0	0.015177
CMCarArrhythm	0.317881	EQ5DMob0	0.020994	BCBMI	0.7132	StudyArm	0.01505
Comorbidities	0.280834	CMCarHyperten	0.017985	ed_le	0.622446	PreviousChemo	0.013022
ROL0	0.276591	PreviousChemo	0.013151	EQ5DMob0	0.60776	PrimaryorMet	0.012692
CMEndDiabetes	0.266086	EQ5DSelCar0	0.011003	CMResAsthma	0.573449	CMPrevMal	0.009341
StudyArm	0.258236	CMEndDiabetes	0.010222	EQ5DVAS0	0.556679	CMCarHyperten	0.00828
EQ5DAnxDep0	0.254403	CMPrevMal	0.009266	PreviousChemo	0.515102	CMResAsthma	0.00717
CMStomInflamm	0.24441	CMResAsthma	0.008121	ROL0	0.472313	CMResCOPD	0.005677
EQ5DMob0	0.236114	CMEndHypo	0.006361	CMEndHyperth	0.433106	CMCarMI	0.005379
CMRheuArth	0.224756	CMSubstAlcohol	0.003922	CMPrevMal	0.405182	CMCarVenous	0.005169
CMCarAngina	0.176192	CMCarMI	0.003742	EQ5DAnxDep0	0.362858	CMEndHypo	0.004344
EQ5DVAS0	0.164569	CMRheuArth	0.003689	StudyArm	0.336571	CMEndDiabetes	0.004095
CMRheuCTD	0.150602	CMCarVenous	0.003277	SWB_overall0	0.327688	CMCarArrhythm	0.002829
DCEmployment	0.142009	CMResCOPD	0.002494	Comorbidities	0.162286	CMCarAngina	0.00265
CMSubstAlcohol	0.136974	CMStomInflamm	0.002476	DCMarital	0.153898	CMRheuArth	0.002563
DCMarital	0.115902	CMRheuLupus	0.002339	PrimaryorMet	0.135378	CMSubstAlcohol	0.002274
CMNeuParkins	0.104832	CMGasPancreas	0.002232	CMRenEndStage	0.11648	CMStomInflamm	0.001276
CMNeuStroke	0.097995	CMCarAngina	0.001642	CMCarMI	0.109787	CMRheuLupus	0.000868
PrimaryorMet	0.096108	CMRenEndStage	0.001634	CMEndDiabetes	0.098698	CMEndHyperth	0.000723
PreviousChemo	0.082822	CMResChronBron	0.000892	CMStomUlcers	0.091996	CMRheuCTD	0.00058
SWB_overall0	0.071112	CMNeuStroke	0.000826	CMGasPancreas	0.084078	CMRenEndStage	0.000444
CMResChronBron	0.063813	CMCarArrhythm	0.000588	EQ5DSelCar0	0.076894	CMCarHeartFail	0.000305
ed_le	0.062876	CMEndHyperth	0.00042	DCEmployment	0.063975	CMNeuParkins	0.000191
CMEndHyperth	0.058035	CMRheuCTD	0.000215	CMCarHyperten	0.04455	CMResChronBron	0.000177
PWB_overall0	0.025423	CMNeuParkins	0.000096	EQ5DPain0	0.037622	CMNeuStroke	0.000131
CMGasChronHep	0	CMCarHeartFail	0	DiseaseSite	0.027685	CMStomUlcers	0.000052
CMCarHeartFail	0	CMGasChronHep	0	CMResChronBron	0.000317	CMGasCirrhosis	0
CMResEmphys	0	CMResEmphys	0	CMSubstDrugs	0	CMResEmphys	0
CMRheuPolymyo	0	CMSubstDrugs	0	CMGasCirrhosis	0	CMSubstDrugs	0
CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMGasChronHep	0	CMGasPancreas	0
CMStomUlcers	0	CMGasCirrhosis	0	CMRheuRhPolymy	0	CMRheuPolymyo	0
CMStomMalabsor	0	CMNeuMyasth	0	CMResEmphys	0	CMNeuMyasth	0
CMSubstDrugs	0	CMNeuMS	0	CMRheuPolymyo	0	CMNeuMS	0
CMNeuMS	0	CMStomMalabsor	0	CMStomMalabsor	0	CMGasChronHep	0
CMNeuMyasth	0	CMStomUlcers	0	CMNeuMyasth	0	CMStomMalabsor	0
CMGasCirrhosis	0	CMRheuPolymyo	0	CMNeuMS	0	CMRheuRhPolymy	0
CMRheuPolymyo	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor	0

Table 8. Feature importance ranks for mobility changes at 18 weeks prediction models LR and RF.

Mobility						
Deterioration			Improvement			
Logistic regression		Random forest	Logistic regression		Random forest	
CMResChronBron	5.681567	FWB_overall0	0.1055	CMEndHyperth	5.327951	EQ5DMob0
CMGasPancreas	4.960102	SWB_overall0	0.091013	EQ5DMob0	4.080502	BCBMI
CMCarHeartFail	4.945931	PWB_overall0	0.077668	CMCarAngina	3.690014	FWB_overall0
CMCarArrhythm	4.130466	EQ5DVAS0	0.073442	CMResCOPD	3.684776	DiseaseSite
CMCarMI	3.865736	DiseaseSite	0.051927	CMCarVenous	2.94743	SWB_overall0
CMSubstAlcohol	3.775084	DCEmployment	0.048892	CMResChronBron	2.584025	PWB_overall0
CMEndHyperth	3.298455	ed_lev	0.040528	CMNeuParkins	2.409078	EQ5DPain0
CMResCOPD	3.033808	EQ5DAnxDep0	0.040475	CMCarHyperten	2.35333	EQ5DVAS0
CMRheuArth	2.903339	Comorbidities	0.034333	CMResAsthma	2.132524	DCMarital
CMRheuLupus	2.42351	DCMarital	0.034039	PreviousChemo	2.022009	EQ5DUsuAct0
CMCarVenous	1.366612	EQ5DUsuAct0	0.033967	CMCarMI	1.74107	DCEmployment
CMEndDiabetes	1.287484	EQ5DPain0	0.033657	CMEndDiabetes	1.617783	Comorbidities
CMResAsthma	1.202195	StudyArm	0.020932	CMEndHypohty	1.571741	EQ5DAnxDep0
PreviousChemo	1.0179	PrimaryorMet	0.019977	Comorbidities	1.328023	ed_lev
CMNeuParkins	0.761727	EQ5DMob0	0.018281	DiseaseSite	1.312387	CMCarHyperten
BCBMI	0.651666	CMResAsthma	0.013156	SWB_overall0	1.307746	PrimaryorMet
DiseaseSite	0.57739	PreviousChemo	0.013074	BCBMI	1.165174	PreviousChemo
CMPrevMal	0.561657	CMCarHyperten	0.011887	CMCarArrhythm	1.071283	StudyArm
PrimaryorMet	0.525226	CMPrevMal	0.010103	CMPrevMal	0.984691	EQ5DSelCar0
CMStomInflamm	0.514437	EQ5DSelCar0	0.009966	CMNeuStroke	0.911514	CMPrevMal
Comorbidities	0.489691	CMCarArrhythm	0.009579	EQ5DUsuAct0	0.856713	CMEndDiabetes
EQ5DPain0	0.463356	CMEndDiabetes	0.009069	FWB_overall0	0.649634	CMCarMI
EQ5DMob0	0.433626	CMCarMI	0.008925	CMRheuArth	0.530132	CMResAsthma
CMNeuStroke	0.422049	CMRheuArth	0.008617	DCMarital	0.480517	CMEndHypohty
EQ5DAnxDep0	0.41443	CMEndHypohty	0.008269	PrimaryorMet	0.457481	CMCarAngina
EQ5DSelCar0	0.396671	CMResCOPD	0.003547	StudyArm	0.292456	CMEndHyperth
EQ5DVAS0	0.283777	CMCarVenous	0.003416	CMStomInflamm	0.249998	CMCarArrhythm
ed_lev	0.263051	CMRheuLupus	0.00247	EQ5DSelCar0	0.224756	CMCarVenous
CMCarHyperten	0.117272	CMStomInflamm	0.002402	CMRheuLupus	0.161346	CMResCOPD
CMEndHypohty	0.10792	CMNeuStroke	0.002397	CMSubstAlcohol	0.118481	CMRheuArth
CMCarAngina	0.087238	CMGasPancreas	0.001793	CMGasPancreas	0.10859	CMResChronBron
FWB_overall0	0.0783	CMCarHeartFail	0.001535	CMRenEndStage	0.1067	CMStomInflamm
EQ5DUsuAct0	0.073808	CMEndHyperth	0.001213	EQ5DVAS0	0.095971	CMNeuParkins
DCMarital	0.057833	CMSubstAlcohol	0.001119	EQ5DAnxDep0	0.092543	CMSubstAlcohol
PWB_overall0	0.057653	CMCarAngina	0.000915	ed_lev	0.087568	CMNeuStroke
StudyArm	0.048044	CMResChronBron	0.000548	EQ5DPain0	0.075561	CMRheuLupus
SWB_overall0	0.039861	CMNeuParkins	0.000193	DCEmployment	0.074864	CMGasPancreas
DCEmployment	0.030947	CMResEmphys	0	PWB_overall0	0.00146	CMCarHeartFail
CMResEmphys	0	CMSubstDrugs	0	CMRheuCTD	0.000971	CMStomMalabsor
CMGasChronHep	0	CMRheuRhPolymy	0	CMCarHeartFail	0.000414	CMSubstDrugs
CMGasCirrhosis	0	CMGasChronHep	0	CMStomUlcers	0	CMGasCirrhosis
CMStomUlcers	0	CMRheuCTD	0	CMNeuMyasth	0	CMGasChronHep
CMRheuPolymy	0	CMNeuMyasth	0	CMStomMalabsor	0	CMRheuRhPolymy
CMRheuRhPolymy	0	CMNeuMS	0	CMGasChronHep	0	CMRheuPolymy
CMRenEndStage	0	CMRenEndStage	0	CMGasCirrhosis	0	CMRheuCTD
CMNeuMS	0	CMStomMalabsor	0	CMResEmphys	0	CMResEmphys
CMSubstDrugs	0	CMStomUlcers	0	CMRheuRhPolymy	0	CMNeuMyasth
CMNeuMyasth	0	CMGasCirrhosis	0	CMRheuPolymy	0	CMNeuMS
CMRheuCTD	0	CMRheuPolymy	0	CMNeuMS	0	CMRenEndStage
CMStomMalabsor	0			CMSubstDrugs	0	CMStomUlcers
CMRheuPolymy	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor

8.5 Paper 5

Time-Series Patient-Reported Data and LSTM Predicting Hospital Utilisation During Chemotherapy

Zuzanna Wójcik

*School of Computer Science
University of Leeds
Leeds, United Kingdom
ORCID: 0009-0007-6214-7736
sczw@leeds.ac.uk*

Vania Dimitrova

*School of Computer Science
University of Leeds
Leeds, United Kingdom
ORCID: 0000-0002-7001-0891
v.g.dimitrova@leeds.ac.uk*

Lorraine Warrington

*Leeds Institute of Medical Research
University of Leeds
Leeds, United Kingdom
ORCID: 0000-0002-8389-6134
l.warrington@leeds.ac.uk*

Galina Velikova

*Leeds Institute of Medical Research
University of Leeds
Leeds, United Kingdom
ORCID: 0000-0003-1899-5942
g.velikova@leeds.ac.uk*

Kate Absolom

*Leeds Institute of Medical Research
University of Leeds
Leeds, United Kingdom
ORCID: 0000-0002-5477-6643
k.l.absolom@leeds.ac.uk*

Samuel D Relton

*Leeds Institute of Health Sciences
University of Leeds
Leeds, United Kingdom
ORCID: 0000-0003-0634-4587
s.d.relton@leeds.ac.uk*

Abstract—Adverse effects of chemotherapy often require acute hospital admissions, which can negatively impact patients' well-being and increase healthcare burden. Identifying the risk of hospital utilisation could support prevention of patients' deterioration and alert medical teams about potential admissions. This study uses patients' clinical and demographic data, patient-reported outcome measures, and time-series symptom severity reports to predict the risk of hospital admissions and triage events within 14 days from completed symptom severity report. Hospital utilisation at any time during chemotherapy was also predicted. The performance of long-short term memory (LSTM) and extreme gradient boosting (XGBoost) models was compared. Nested cross-validation enabled robust hyperparameter tuning and model evaluation using unseen data. Patient representatives and a clinical oncologist were consulted during the study design to support its clinical relevance. LSTM outperformed XGBoost at short-term predictions of hospital admission (balanced accuracy=0.780, AUC=0.845) and triage (balanced accuracy=0.706, AUC=0.779). However, XGBoost performed better at long-term predictions. The results suggest that LSTM processed complex data with sudden fluctuations better than XGBoost. However, classical ML might be sufficient for longer-term outcome predictions. If further explored, these models could prompt hospital contact prior to patient's deterioration and prevent admission or alert medical team of potential hospitalisation.

Index Terms—patient-reported data, time-series data, LSTM, hospital utilisation predictions

I. INTRODUCTION

Chemotherapy toxicity is associated with physical, psychological, and psychosocial symptoms negatively impacting patients' quality of life (QoL). Adverse effects of chemotherapy can lead to initial or prolonged hospitalisation, permanent disability, life-threatening situations, or death [1]. Unplanned healthcare utilisation during chemotherapy can be detrimental

to healthcare quality, through increased burden on hospitals. Identifying symptoms that contribute to acute presentations to hospitals of cancer patients can support planning for emergency admissions, increase the quality of care, and reduce healthcare costs [2]. Predictions of acute admissions may help to prompt hospital contact, activate emergency alerts to the medical team, and prevent patient deterioration [3].

Understanding symptom experiences and severity is possible due to the use of patient-reported data collected throughout chemotherapy. Questionnaires, such as symptom severity reports or patient-reported outcome measures (PROMs) can capture patients' perspective on their own health status [3]. These have been useful in clinical practice through increasing patient involvement in shared treatment decision making. There is also existing evidence that patient-reported data can add predictive value to machine learning (ML) models [4], [5]. However, there is limited research on deep learning (DL) approaches applied on such data [6].

To explore the potential of DL models applied to time-series patient-reported data, this study investigates the performance of long-short term memory (LSTM) model, a type of recurrent neural network (RNN), in the short-term predictions of acute hospital admissions and triage events occurring within 14 days from completing a symptom severity report. Predictions of admissions and clinical triage events following a report at any time during chemotherapy are also explored. All models use time-series symptom severity reports and PROMs, as well as static data, such as clinical and demographic information. LSTM performances are compared with extreme gradient boosting (XGBoost) [7], which acts as a baseline in this study.

A. Related work

Hospital utilisation is a common outcome in medical prediction models [8], suggesting that AI has a potential to support hospitals in improving healthcare quality. Hospital utilisation predictions can involve detecting changes in data happening over a long period of time, or sudden events. According to a literature review conducted by Teo et al., [8] classical statistical and ML models can perform well in predicting outcomes not limited in time (e.g., cardiovascular disease readmission). However, the LSTM model was superior in predicting outcomes that required capturing sudden changes in patterns of data (e.g., intensive care unit readmissions).

Studies predicting hospital utilisation often use statistical models, or classical ML methods [9], [10] with limited capability to capture complex dependencies and patterns in multi-dimensional or time-series data [11]. There are existing studies exploring DL models for predictions of oncology outcomes, but the main focus has been on toxicity symptoms [12]–[14], rather than hospital utilisation. These studies have used LSTM model, which is the most common approach to process time-series data in healthcare [11]. Of these studies, only Wang et al., in a paper predicting head and neck cancer symptoms [12], and its follow-up study [13] used patient-reported data. Both studies described a common challenge of using longitudinally collected patient-reported data, which is the irregularity in patients' reports. Therefore, they explored missing data imputation techniques. Nevertheless, these techniques might introduce bias and prevent the model from learning from the patterns of missingness, which could hold information on patients' health status.

Furthermore, Wang et al., [13] have only used patient-reported data without the inclusion of clinical or demographic information. Combining time-series data with static data (e.g., demographic) adds a challenge in DL model development. A simple approach is a repetition of the same static information for each time-point generated by the same patient [11]. However, there are possibilities to use the DL network architecture to handle static data inclusion more efficiently. Time-series and static data can be processed by different network layers as separate inputs. This is the most common approach in medical outcome predictions using static and temporal data [11]. For example, Li et al., [15] used a multi-modal fusion approach that uses the representation of static data as the initial hidden state of RNN. Another study put static input separately through a fully connected layer, and time-series input through LSTM and convolutional neural network layers, combining both outputs at the end [16].

Cancer prediction and diagnosis models rarely include patient-reported data [6], missing the information on patients' perspective on their health. Even though such studies predict clinical outcomes, there is also no evidence of patient or clinician engagement during the study design. Since the development of AI models is not based on patients' perspective or clinical knowledge, it might not actually serve its purpose or address needs of diverse group of people in an equal and

fair way. Involving patients and public input is essential not only for the feasibility of novel models, but also for increasing public trust in AI research [17].

B. Contributions

To our knowledge, there is no existing research predicting chemotherapy toxicity related hospital utilisation using DL models and time-series patient-reported data combined with static clinical and demographic information. The proposed approach uses LSTM models to predict admissions and triage events within 14 days from the completion of a symptom severity report. A successful system that can alert to potential complications during chemotherapy can be extremely useful to prevent serious complications or hospitalisation [3]. Therefore, the main contribution of this study is the short-term prediction of hospital utilisation based on longitudinal patient-reported symptom severity reports using LSTM models. Admissions and triage events happening at any time of the chemotherapy were also predicted. In this study the information from missing reports was used, and not imputed, providing the model with an opportunity to learn from the irregularities in reporting. This approach was decided following the consultations with patient representatives and a clinical oncologist during the study design. They suggested that the irregularity of reporting might provide the information of patients' health status (e.g., feeling too fatigued to report).

II. METHODS

A. Overall methodology

This study explores the performance of LSTM and XGBoost using clinical, demographic, and longitudinally collected patient-reported data predicting admissions and triage happening within 14 days from a completed symptom severity report (short-term predictions) and at any time during chemotherapy (long-term predictions). Figure 1 illustrates the overall methodology of this research. LSTM is the most common model to process time-series data in healthcare [11], so was selected as a DL model in this study. XGBoost is a baseline model, as it is often compared to DL models when predicting outcomes from numerical, tabular data [7]. XGBoost also performed very well in previous studies predicting hospital utilisation, outperforming other ML models [18]. Patient representatives were consulted during the study design to provide the patients' perspective on the subject. The decisions regarding outcomes and feature engineering were influenced by these consultations. Furthermore, a clinical oncologist was involved, which improved the robustness of the design and its relevance to clinical practice.

B. Dataset

The data were collected from 256 patients initiating systemic treatment for colorectal, breast, or gynecologic cancers in an eRAPID clinical trial [3] at Leeds Cancer Centre (United Kingdom) between January 22, 2015, and June 11, 2018.

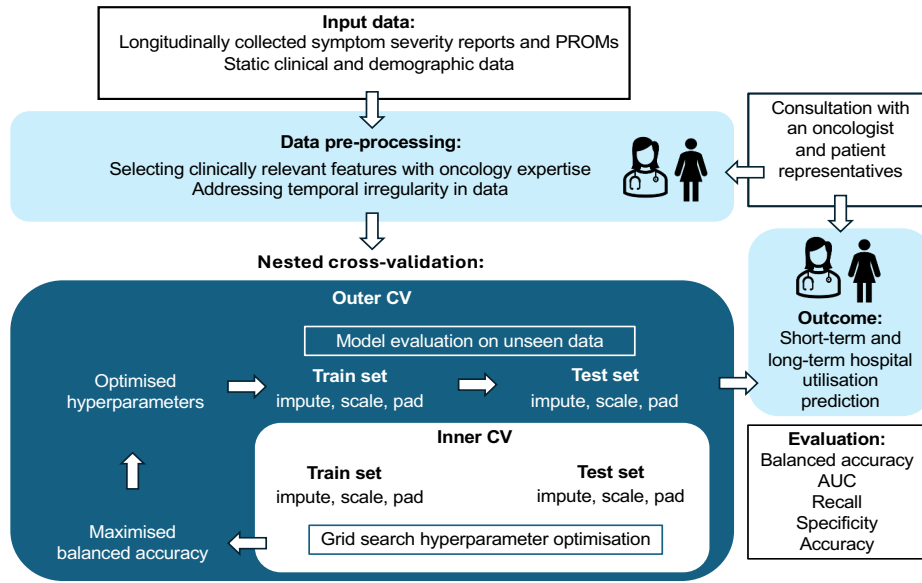


Fig. 1. Flow diagram illustrating data flow and the nested cross-validation (CV) pipeline for implementing LSTM and XGBoost models in this study.

The static data are clinical and demographic information derived from electronic healthcare records. Time-series features include symptom severity reports and PROMs. The patients were asked to complete symptom severity reports weekly for 18 weeks. There were 3260 reports in total. Each included multiple symptoms (pain, nausea, vomiting, diarrhoea, temperature, chills, physical ability, appetite, fatigue, sore mouth, indigestion, shortness of breath) and a scale for assessing severity from 0 (no symptom) to 3 (severe). PROMs were completed by participants at 4 time-points (baseline at the start of the trial, and at 6-, 12-, and 18-week follow-up). For each time-point, 6 variables were from Five-dimensional Visual Analogue Scale (EQ-5D-VAS), including mobility, self-care, usual activities, pain/discomfort, anxiety/depression, and self-rated health status; 4 variables from Functional Assessment of Cancer Therapy - General 28 items (FACT-G), including aggregated scores of physical, social, emotional and functional well-being; and 2 remaining variables were social and role scale from EORTC Core QoL Questionnaire (QLQ-C30). See eRAPID clinical trial report [3] for details on questionnaires.

C. Outcome variables

The outcomes of this study are the short-term and long-term predictions of hospital utilisation:

Short-term predictions of hospital utilisation:

- Admission within 14 days (whether a patient was admitted to hospital within 14 days from completing the report)
- Clinical triage within 14 days (whether a patient contacted hospital within 14 days from completing the report)

Long-term predictions of hospital utilisation:

- Admission (whether a patient was admitted to hospital at any time during chemotherapy, following the report)
- Clinical triage (whether a patient contacted hospital at any time during chemotherapy, following the report)

The 14-day window of admission was chosen by a clinical oncologist as a useful time frame for detecting symptom deterioration in clinical practice, which would enable preparing for or avoiding potential admission. These outcomes were computed from the dates of admissions and triage episodes during the clinical trial. A binary column was created to indicate if the event happened (1) or not (0), following the completion of the symptom report. The long-term predictions don't limit the time frame to any specific amount of days, so the patterns in data can be less complex to capture [8].

D. Data pre-processing

1) *Feature engineering*: Since symptom severity reports and PROMs were collected at different time-steps, a PROMs value was allocated to each report depending on the time-step of completion (for example, for each report completed between baseline and 6 weeks, the PROMs collected at baseline were

allocated; for reports between 6 and 12 weeks, the PROMs collected at 6 weeks were allocated). The static data were kept separately during the data pre-processing stage for LSTM. However, for XGBoost, to each report, repeated values of patient's static data were added. From all available features, the selection of clinically relevant ones was performed by a clinical oncologist.

2) *Addressing irregularity in data:* LSTM model requires all data sequences to have the same length for each subject. However, the number of reports submitted by each patient varied due to missed or frequent reporting. The challenge of irregularity can be addressed in 2 different ways: 1) by aggregating the reports that were completed more than once a week (e.g., taking average or the lowest score out of all reports completed during that week) and imputing the missing reports [12], or 2) by leaving the irregularity in data, padding to longest sequence, and including a variable which indicates the time interval between the current and previous report. Due to a significant variability in frequency of reporting, option 1) would create too much synthetic data and would prevent the model from learning on real data. Therefore, a variable was created to indicate the number of days since the previous report. It provides the opportunity for the model to learn from the temporal pattern of reporting. Missing reports were not imputed, but k-nearest neighbors (KNN, $k=5$) was used to impute any sporadic missing data. There are very few cases of missingness in this dataset. Three patients were outliers, and therefore their data were removed. The longest sequence (43 reports) was then selected and all sequences were padded (filled with a particular value) to match its length within the nested cross-validation loops. This is a common approach to adjust the size of the data in DL model development [19].

E. Model implementation pipeline

1) *LSTM model architecture:* The model accepts 2 separate inputs: static data (demographic and clinical information) and time-series data (symptom severity reports and PROMs). The LSTM layer processes time-series data, and outputs the hidden representation of time-series sequences. The fully connected layer receives the concatenated input of static data and the LSTM output (processed temporal dependencies). The output is converted by sigmoid activation to a probability for binary classification. Therefore, final output is the binary prediction of event not happening (0) or happening (1). The model architecture is presented in Figure 2. The hidden size, number of layers, learning rate were established through hyperparameter tuning. Due to large class imbalance in data (out of 3260 reports, there were: 640 admissions, 175 admission within 14 days, 1212 triage events, 454 triage events within 14 days), class weight factor was set to 1.

2) *Nested cross-validation:* To ensure robust hyperparameter optimisation, prevent data leakage and over-estimation of results, the models were implemented through the loops of nested cross-validation [20] (Figure 1). The hyperparameter

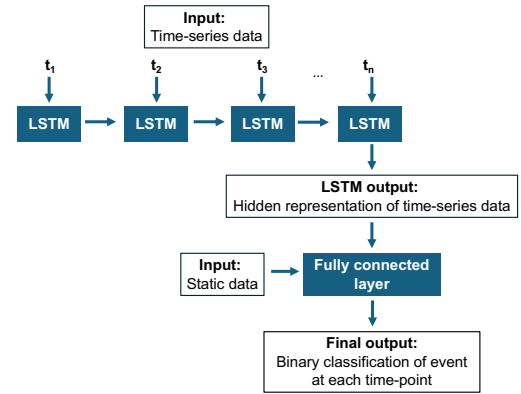


Fig. 2. Flow diagram illustrating the LSTM model. Time-series input is processed through LSTM layer, which output is then passed through fully connected layer, receiving separate static input. The model hyperparameters were established during nested cross-validation process.

tuning was performed in the inner cross-validation loop and in outer cross-validation loop model was evaluated on unseen test set. In each loop a stratified split of patients to train and test sets was completed. The continuous variables were scaled to unit variance, and missing values were imputed with KNN($k=5$). The sequences were padded with -999 in order to avoid confusion with zeros existing in the dataset. To further avoid model learning from padded data, masking was applied on padded values. All these steps were performed within loops to avoid data leakage to test sets.

The hyperparameter tuning in the inner loop involved the following hyperparameter sets: hidden size: [64, 128], number of layers: [1, 2], and learning rate: [0.001, 0.0001]. The number of inner folds was 2. Each iteration involved 20 epochs, but early stopping was introduced if the validation loss was not improved in 5 epochs. In the outer loop, the model with hyperparameters selected in the inner loop was trained during 50 epochs and evaluated based on the average of the performance metrics over 3 iterations. The model was trained with "BCEWithLogitsLoss" loss function combining sigmoid activation and binary cross-entropy loss, due to the required binary output (event happening or not happening).

The pipeline for XGBoost was the same, with some adjustments. In each loop, the training set was balanced using Synthetic Minority Oversampling Technique (SMOTE), to help the model learn from the originally imbalanced data. This step was unnecessary in LSTM, as the class weight factor was adjusted. Furthermore, XGBoost does not require as much running time as LSTM, so the nested cross-validation included 5 iterations of both inner and outer loops. In LSTM, to support code efficiency, the number of iterations was limited [7]. Another difference between the 2 models was combining

static and temporal data. XGBoost cannot take 2 separate inputs, so the static data was repeated for each time-step. The hyperparameters grid for XGBoost involved: number of estimators: [50, 100], learning rate: [0.01, 0.05], maximum depth: [3, 4], subsample: [0.7, 0.8], and percentage of features used for building each tree: [0.7, 0.8].

F. Evaluation metrics

The models were evaluated with 5 different performance metrics. Balanced accuracy was selected as the main one, as it is useful at assessing performance of models serving as a screening system, being able to capture model's bias towards one class. Area under the ROC curve (AUC) was also used, as it is a very common metric in ML studies. Even though it is not recommended for imbalanced data, it can be used for between-studies comparison. Recall and specificity are considered as important metrics which explain the model's abilities to discriminate between classes. Plotted together, they can help visualise if the model is struggling with a positive or negative class. Accuracy was also reported due to its common use in research, but is not discussed, as it is not relevant for models tested on imbalanced data.

III. RESULTS

A. Short-term predictions

Table I and Figure 3 present all results for short-term predictions of admission and triage. Overall, LSTM performed better than XGBoost. Balanced accuracy indicated advantage of using LSTM (0.780 for admission within 14 days and 0.706 for triage within 14 days) over XGBoost (0.545 for admission within 14 days and 0.532 for triage within 14 days) for short-term predictions. AUC of 0.845 suggests a very good performance of LSTM for short-term admissions prediction, which is also higher than the AUC of XGBoost (0.724). LSTM for short-term predictions of triage also resulted with higher AUC (0.779) than XGBoost (0.687). Specificity of LSTM predicting admission within 14 days is high (0.846), with lower recall (0.715), suggesting that the model is slightly better at detecting negative class. However, the overall performance is balanced, as all metrics are considerably high. The specificity of XGBoost (0.991), which means that the model is very good at detecting negative class. However, it has an extremely low recall of 0.098, suggesting that the model tends to predict no admissions within 14 days from most reports, generating a lot of false negative results. For triage, XGBoost has a similar bias towards positive class, with recall of 0.085 and specificity of 0.979. LSTM predicted triage within 14 days with much higher recall (0.699), and lower specificity (0.712). Overall, the results of LSTM are much more balanced than of XGBoost, without bias towards any class.

Overall, short-term predictions of admissions had a higher performance than of triage. However, balanced accuracy was only slightly higher for admission within 14 days than triage within 14 days for both LSTM and XGBoost, with the difference lower than 0.1. All other metrics consistently suggest better performance of both models for short-term predictions

TABLE I
LSTM AND XGBOOST PERFORMANCE OF SHORT-TERM AND LONG-TERM PREDICTIONS OF HOSPITAL UTILISATION.

Evaluation metric	Model	Admission within 14 days	Triage within 14 days	Admission	Triage
Balanced accuracy	LSTM	0.780	0.706	0.717	0.776
	XGBoost	0.545	0.532	0.813	0.836
AUC	LSTM	0.845	0.779	0.775	0.839
	XGBoost	0.724	0.687	0.939	0.917
Recall	LSTM	0.715	0.699	0.618	0.823
	XGBoost	0.098	0.085	0.654	0.795
Specificity	LSTM	0.846	0.712	0.816	0.729
	XGBoost	0.991	0.979	0.973	0.877
Accuracy	LSTM	0.836	0.709	0.773	0.765
	XGBoost	0.943	0.853	0.914	0.912

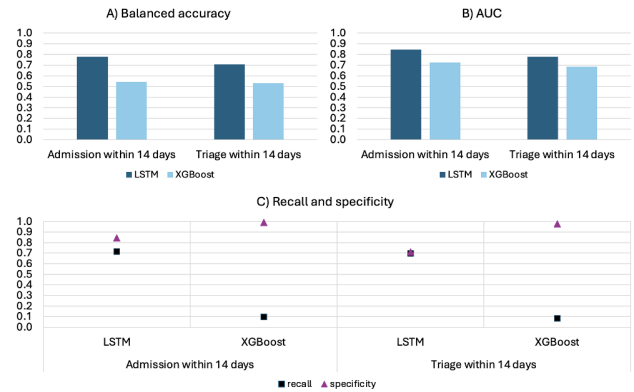


Fig. 3. Plotted balanced accuracy, AUC, recall and specificity of LSTM and XGBoost predicting admission and triage within 14 days.

of admissions than triage. It suggest that predicting admission within 14 days led to fewer false negative and false positive results than predicting triage within 14 days (Figure 3).

B. Long-term predictions

Figure 4 and Table I present the results of long-term hospital utilisation predictions. Overall, the long-term predictions were good for both models. However, unlike in short-term predictions, XGBoost performed better than LSTM, achieving higher balanced accuracy, AUC and specificity for both admission and triage predictions. Recall was very similar for both models. XGBoost resulted with higher recall for admissions prediction, and LSTM for triage prediction. The balanced accuracy of 0.813 and AUC of 0.939 suggest that XGBoost can successfully predict admission during chemotherapy with a good performance. Similarly, balanced accuracy of 0.836 and AUC of 0.917 suggest a good performance of XGBoost predicting clinical triage. The balanced accuracy for LSTM was 0.717 (admission prediction), and 0.776 (triage prediction) which is only slightly lower than for XGBoost. The recall in predicting admission was moderate for both models (between 0.6 and 0.7), suggesting a bias towards a negative class. Models predicting triage had a higher performance than predicting admissions, based on balanced accuracy. However, the difference was lower than 0.1 for both models.

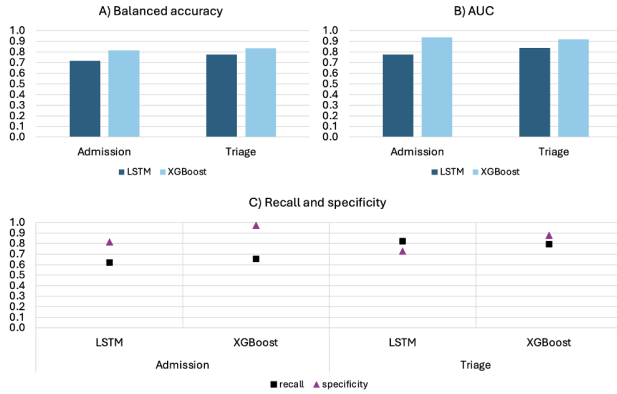


Fig. 4. Plotted balanced accuracy, AUC, recall and specificity of LSTM and XGBoost predicting admission and triage.

IV. DISCUSSION

The results of this study indicate that LSTM model is successful at short-term predictions of hospital utilisation and its performance is superior to XGBoost. However, LSTM performed slightly worse than XGBoost at predicting hospital utilisation happening at any time during chemotherapy.

LSTM performing better than XGBoost in short-term hospital predictions is consistent with literature suggesting that DL models are able to process more complex and sudden fluctuations in the dataset [8]. Therefore, in time-series scenarios where the outcomes happen shortly after the input time-step, deep network architectures, such as LSTM, have more potential to succeed. On the other hand, the performance of long-term predictions of admission and triage was better for XGBoost than LSTM. However, the difference in these performances was not as high as for short-term predictions. There is previous evidence that XGBoost often outperforms DL models when predicting outcomes from numerical, tabular data [7]. The complexity of sudden outcome following report was removed in these scenario, which might have strengthen the performance of XGBoost. Another explanation for the differences in the model performances can be derived from significance testing of the differences between variables for each class of the outcomes. Table II presents the number of variables which are statistically significantly different for classes in each outcome. It indicates that more features differed significantly for classes in admission and triage than for classes in admission and triage within 14 days. These differences might have provided a more structured discriminatory advantage for XGBoost in predicting admission and triage. For short-term predictions more complex dependencies in data were necessary to process, which LSTM succeeded in.

The overall performance of all models is good, which can be compared with other studies. Peterson et al., [10], predicted emergency department visits and hospital admissions during chemotherapy from clinical, demographic and patient-reported data using ML models. The reported AUC for the best model was 0.783, which is lower than the highest AUC of LSTM

TABLE II
NUMBER OF FEATURES WITHIN DIFFERENT SIGNIFICANCE LEVELS OF THE DIFFERENCE BETWEEN CLASSES FOR EACH OUTCOME.

	Admission within 14 days	Triage within 14 days	Admission	Triage
$p < 0.001$	14	16	20	21
$p < 0.01$	6	6	8	8
$p < 0.05$	5	6	1	4
$p > 0.05$	18	15	14	10

for short-term admission predictions (0.845), and XGBoost for admission during chemotherapy predictions (0.939). In another study predicting emergency visits and hospital admissions during cancer treatment [9], the AUC of gradient tree boosting was 0.798, which is also lower, compared to the results in this study. Good performance of these models might be the result of robust design with nested cross-validation, as this approach can reduce bias in model selection [21].

The strengths of this study include the prediction of an outcome within the short time frame of patient-reported symptom severity. If applied in clinical practice, such system could help guide the identification of patient deterioration and support to prevent hospital admission. The involvement of patient representatives and a clinician in a study design provided perspectives of stakeholders which is necessary for any AI system to be included in clinical practice. Furthermore, the model implementation with the nested cross-validation pipeline enabled robust hyperparameter tuning and minimised the risk of data leakage [20]. Additionally, multiple evaluation metrics reported enable in-depth explanation of model performance and between-study comparison.

The interpretation of the study should also consider its limitations. The inherent bias of participant samples recruited to clinical trials could affect the representability of the sample. Therefore, future work should validate the models on an external dataset. The sample size in this study is also relatively small, which could have negatively affected the performance of LSTM, as DL models require more data than classical ML models [6]. Furthermore, even though LSTM and XGBoost were implemented through the same pipeline, the differences included the combination of static and time-series features, and handling class imbalance. The discrepancies could have affected the performance of the models, but were unavoidable due to different model architectures. The number of iterations was also inconsistent between the 2 models, which could have prevented LSTM from learning effectively. In the future, the trade-off between model efficiency and the robustness of the design should be explored. Future work should also involve more structured focus groups with patients and clinicians, including the discussion around results interpretation.

V. CONCLUSION

This study used time-series patient reported data, as well as static demographic and clinical information to predict hospital utilisation during chemotherapy. LSTM model, by detecting sudden changes in data, successfully predicted admissions and

triage events happening within 14 days from symptom severity reports, outperforming XGBoost. However, admissions and triage predictions at any time during chemotherapy resulted with higher performance of XGBoost. It suggests that simple ML models might be better for less complex predictions. If further explored, this study could support reducing acute hospital utilisation during chemotherapy by detecting risk of deterioration in a timely manner. Furthermore, it encourages patient and clinician involvement in AI research and supports informed shared treatment decision-making.

VI. ACKNOWLEDGEMENTS

This research is supported in part by UK Research and Innovation [EP/S024336/1]. Ethics approval was granted by School of Medicine Research Ethics Committee [MREC 22-098], University of Leeds. The authors thank patient representatives from Patient Centred Outcomes Research (PCOR) in the Leeds Institute of Medical Research for consultation, and the patients and clinicians involved in the eRAPID clinical trial for providing the data. The authors also thank Allan Pang and Alexander Coles for the help with study design.

REFERENCES

- [1] I. Ingrand, G. Defosse, C. Lafay-Chebassier, F. Chavant, A. Ferru, P. Ingrand, and M.-C. Pérault-Pochat, "Serious adverse effects occurring after chemotherapy: A general cancer registry-based incidence survey," *British Journal of Clinical Pharmacology*, vol. 86, no. 4, pp. 711–722, 2020, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/bcp.14159>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1111/bcp.14159>
- [2] P. H. Dufon, M. F. Gertz, R. Jarden, and M. Krishnasamy, "Methodological approaches to measuring the incidence of unplanned emergency department presentations by cancer patients receiving systemic anti-cancer therapy: a systematic review," *BMC Medical Research Methodology*, vol. 22, no. 1, p. 75, Dec. 2022. [Online]. Available: <https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01555-3>
- [3] K. Absolom, L. Warrington, E. Hudson, J. Hewison, C. Morris, P. Holch, R. Carter, A. Gibson, M. Holmes, B. Clayton, Z. Rogers, L. McParland, M. Conner, L. Glidewell, B. Woroncow, B. Dawkins, S. Dickinson, C. Hulme, J. Brown, and G. Velikova, "Phase III Randomized Controlled Trial of eRAPID: eHealth Intervention During Chemotherapy," *Journal of Clinical Oncology*, vol. 39, no. 7, pp. 734–747, Mar. 2021, publisher: Wolters Kluwer.
- [4] J.-a. Sim, Y. A. Kim, J. H. Kim, J. M. Lee, M. S. Kim, Y. M. Shim, J. I. Zo, and Y. H. Yun, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Scientific Reports*, vol. 10, p. 10693, Jul. 2020. [Online]. Available: <https://www.nature.com/articles/s41598-020-67604-3>
- [5] Z. Wójcik, V. Dimitrova, L. Warrington, G. Velikova, and K. Absolom, "Using Machine Learning to Predict Unplanned Hospital Utilization and Chemotherapy Management From Patient-Reported Outcome Measures," *JCO Clinical Cancer Informatics*, no. 8, p. e2300264, Apr. 2024, publisher: Wolters Kluwer. [Online]. Available: <https://ascopubs.org/doi/10.1200/CCI.23.00264>
- [6] D. Krepper, M. Cesari, N. J. Hubel, P. Zelger, and M. J. Sztankay, "Machine learning models including patient-reported outcome data in oncology: a systematic literature review and analysis of their reporting quality," *Journal of Patient-Reported Outcomes*, vol. 8, no. 1, p. 126, Nov. 2024. [Online]. Available: <https://doi.org/10.1186/s41687-024-00808-7>
- [7] R. Schwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253521002360>
- [8] K. Teo, C. W. Yong, J. H. Chuah, Y. C. Hum, Y. K. Tee, K. Xia, and K. W. Lai, "Current Trends in Readmission Prediction: An Overview of Approaches," *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 11117–11134, Aug. 2023. [Online]. Available: <https://doi.org/10.1007/s13369-021-06040-5>
- [9] J. C. Hong, D. Niedzwiecki, M. Palta, and J. D. Tenenbaum, "Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm," *JCO Clinical Cancer Informatics*, no. 2, pp. 1–11, Dec. 2018.
- [10] D. J. Peterson, N. P. Ostberg, D. W. Blayney, J. D. Brooks, and T. Hernandez-Boussard, "Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions," *JCO Clinical Cancer Informatics*, no. 5, pp. 1106–1126, Dec. 2021. [Online]. Available: <https://ascopubs.org/doi/10.1200/CCI.21.00116>
- [11] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time Series Prediction Using Deep Learning Methods in Healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–29, Mar. 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3531326>
- [12] Y. Wang, L. Van Dijk, A. S. R. Mohamed, C. D. Fuller, X. Zhang, G. E. Marai, and G. Canahuete, "Predicting late symptoms of head and neck cancer treatment using LSTM and patient reported outcomes," *Proceedings. International Database Engineering and Applications Symposium*, vol. 2021, pp. 273–279, Jul. 2021.
- [13] Y. Wang, L. V. Dijk, A. S. R. Mohamed, M. Naser, C. D. Fuller, X. Zhang, G. E. Marai, and G. Canahuete, "Improving Prediction of Late Symptoms using LSTM and Patient-reported Outcomes for Head and Neck Cancer Patients," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, Jun. 2023, pp. 292–300, iSSN: 2575-2634. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10337216>
- [14] S. Chae, W. N. Street, N. Ramaraju, and S. Gilbertson-White, "Prediction of Cancer Symptom Trajectory Using Longitudinal Electronic Health Record Data and Long Short-Term Memory Neural Network," *JCO Clinical Cancer Informatics*, no. 8, p. e2300039, Mar. 2024, publisher: Wolters Kluwer. [Online]. Available: <https://ascopubs.org/doi/abs/10.1200/CCI.23.00039>
- [15] D. Li, P. Lyons, J. Klaus, B. Gage, M. Kollef, and C. Lu, "Integrating Static and Time-Series Data in Deep Recurrent Models for Oncology Early Warning Systems," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 913–936. [Online]. Available: <https://dl.acm.org/doi/10.1145/3459637.3482441>
- [16] C. Lin, Y. Zhang, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi, "Early Diagnosis and Prediction of Sepsis Shock by Combining Static and Dynamic Information Using Convolutional-LSTM," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, Jun. 2018, pp. 219–228, iSSN: 2575-2634. [Online]. Available: <https://ieeexplore.ieee.org/document/8419365>
- [17] S. Banerjee, P. Alsop, L. Jones, and R. N. Cardinal, "Patient and public involvement to build trust in artificial intelligence: A framework, tools, and case studies," *Patterns*, vol. 3, no. 6, p. 100506, Jun. 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2666389922000988>
- [18] S. Han, T. J. Sohn, B. P. Ng, and C. Park, "Predicting unplanned readmission due to cardiovascular disease in hospitalized patients with cancer: a machine learning approach," *Scientific Reports*, vol. 13, no. 1, p. 13491, Aug. 2023, publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41598-023-40552-4>
- [19] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, and N. D. B. Bruce, "Position, Padding and Predictions: A Deeper Look at Position Information in CNNs," *International Journal of Computer Vision*, vol. 132, no. 9, pp. 3889–3910, Sep. 2024. [Online]. Available: <https://doi.org/10.1007/s11263-024-02069-9>
- [20] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," Jul. 2022, arXiv:2207.07048 [cs]. [Online]. Available: <http://arxiv.org/abs/2207.07048>
- [21] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, p. 91, Feb. 2006. [Online]. Available: <https://doi.org/10.1186/1471-2105-7-91>

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	N/A
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	Section 2.1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	Section 2.1
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	Section 2.1
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	Section 2.2
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	Section 2.2.2
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	Section 2.2.1
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	Section 2.2.1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	Section 2.2.3
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	Section 2.2.4
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	Section 2.2.4
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	N/A
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	Section 2.2.4



St. Michael's
Inspired Care.
Inspiring Science.

1

Figure A1: PRISMA checklist for scoping reviews part 1 (Chapter 2).

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	Figure 2.1
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	Section 2.3.1
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	N/A
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Section 2.3
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	Section 2.3
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	Section 2.4
Limitations	20	Discuss the limitations of the scoping review process.	Section 2.4.7
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	Section 2.4.8
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	Thesis acknowledgements

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. ;169:467–473. doi: 10.7326/M18-0850



St. Michael's
 Inspired Care.
 Inspiring Science.

2

Figure A2: PRISMA checklist for scoping reviews part 2 (Chapter 2).

Table A1: Search strategy for the scoping review (Chapter 2).

Database	Search strategy
Web of Science, IEEE Xplore, ACM Digital Library	“artificial intelligence” OR “AI” OR “machine learning” OR “deep learning” OR “neural network” OR “neural networks” OR algorithm* OR “rule-based” OR “rule based” AND “predictive model” OR “predictive models” OR “predictive modelling” OR “predictive modeling” OR predict* OR forecast* OR prognos* OR “predicting outcomes” AND “symptom tracking” OR “symptom-tracking” OR “symptom tracker” OR “symptom trackers” OR “symptom assessment” OR “symptom assessing” OR “symptom-assessing” OR “symptom report” OR “symptom reports” OR “symptom reporting” OR “symptom-reporting” OR “symptom monitoring” OR “symptom-monitoring” OR “patient-reported symptom” OR “patient-reported symptoms” OR “patient reported symptom” OR “patient reported symptoms” OR “self-reported symptom” OR “self-reported symptoms” OR “self reported symptom” OR “self reported symptoms” OR “patient-reported outcome” OR “patient-reported outcomes” OR “patient reported outcome” OR “patient reported outcomes” OR “patient-reported outcome measure” OR “patient-reported outcome measures” OR “patient reported outcome measure” OR “patient reported outcome measures” OR “PROMs” OR “self-reported outcome” OR “self-reported outcomes” OR “self reported outcome” OR “self reported outcomes” OR “patient-reported side effect” OR “patient reported side effect” OR “patient reported side effects” OR “patient-reported side effects” OR “self-reported side effect” OR “self reported side effect” OR “self reported side effects” OR “self-reported side effects”
Cochrane Central Register of Controlled Trials	“artificial intelligence” OR “AI” OR “machine learning” OR “deep learning” OR “neural network” OR “neural networks” OR algorithm* OR “rule-based” OR “rule based” AND “predictive model” OR “predictive models” OR “predictive modelling” OR “predictive modeling” OR predict* OR forecast* OR prognos* OR “predicting outcomes” AND “symptom tracking” OR “symptom-tracking” OR “symptom tracker” OR “symptom trackers” OR “symptom assessment” OR “symptom assessing” OR “symptom-assessing” OR “symptom report” OR “symptom reports” OR “symptom reporting” OR “symptom-reporting” OR “symptom monitoring” OR “symptom-monitoring” OR “patient reported symptom” OR “patient reported symptoms” OR “self reported symptom” OR “self reported symptoms” OR “patient reported outcome” OR “patient reported outcomes” OR “patient reported outcome measure” OR “patient reported outcome measures” OR “PROMs” OR “self reported outcome” OR “self reported outcomes” OR “patient reported side effect” OR “patient reported side effects” OR “self reported side effect” OR “self reported side effects”
Medline	exp Artificial Intelligence/ OR AI.mp. OR exp Machine Learning/ OR exp Deep Learning/ OR exp Neural Networks, Computer/ OR exp Algorithms/ OR algorithm*.mp. OR “rule-based”.mp. OR “rule based”.mp AND “predictive model”.mp. OR “predictive models”.mp. OR “predictive modelling”.mp. OR “predictive modeling”.mp. OR predict*.mp. OR forecast*.mp. OR prognos*.mp. OR “predicting outcomes”.mp. AND (symptom adj2 (track* or -track* or assess* or -assess* or report* or -report* or monitor* or -monitor*)).mp. OR “patient-reported symptom”.mp. OR “patient-reported symptoms”.mp. OR “self-reported symptom”.mp. OR “self-reported symptoms”.mp. OR exp Patient Reported Outcome Measures/ OR “patient-reported outcome measure”.mp. OR “patient-reported outcome measures”.mp. OR “patient-reported outcome”.mp. OR “patient-reported outcomes”.mp. OR PROMs.mp. OR “self-reported outcome”.mp. OR “self-reported outcomes”.mp. OR “patient-reported side effect”.mp. OR “patient-reported side effects”.mp. OR “self-reported side effect”.mp. OR “self-reported side effects”.mp.
Embase	exp artificial intelligence/ OR AI.mp. OR exp machine learning/ OR exp deep learning/ OR exp artificial neural network/ OR “neural networks”.mp. OR exp algorithm/ OR algorithm*.mp. OR “rule-based”.mp. AND exp predictive model/ OR “predictive models”.mp. OR “predictive modelling”.mp. OR “predictive modeling”.mp. OR predict*.mp. OR forecast*.mp. OR prognos*.mp. OR “predicting outcomes”.mp. AND (symptom adj2 (track* or -track* or assess* or -assess* or report* or -report* or monitor* or -monitor*)).mp. OR “patient-reported symptom”.mp. OR “patient-reported symptoms”.mp. OR “self-reported symptom”.mp. OR “self-reported symptoms”.mp. OR exp patient reported outcome/ OR “patient-reported outcomes”.mp. OR “self-reported outcome”.mp. OR “self-reported outcomes”.mp. OR “patient-reported side effect”.mp. OR “patient-reported side effects”.mp. OR “patient-reported outcome measure”.mp. OR “patient-reported outcome measures”.mp. OR PROMs.mp. OR “self-reported side effect”.mp. OR “self-reported side effects”.mp.

Table A2: Study characteristics and pre-processing methods used by studies included in the review (Chapter 2).

Authors	Year	Health domain	Sample size	Was outcome self-reported?	Imbalanced data acknowledged?	Need for balancing reported?	Whats the class distribution?	Upsampling type	Missingness reported?
Klent et al [91]	2022	orthopedics	4526	yes	no	no	not reported	n/a	no
Sandham et al [165]	2022	palliative care	751	no	yes	no	n/a	imbalanced	yes
Tennenhouse et al [92]	2020	mental health	637	no	no	no	imbalanced	n/a	yes
Zhang et al [93]	2021	orthopedics	1508	yes	yes	yes	n/a	partially balanced	yes
Zhang et al [94]	2022	orthopedics	2840	yes	yes	yes	n/a	partially balanced	yes
Haeberle et al [162]	2021	orthopedics	3147	no	yes	yes	n/a	testing data	yes
Verna et al [95]	2021	orthopedics	218	yes	no	no	no classification	n/a	no
Buus et al [96]	2022	orthopedics	201	yes	no	no	balanced	n/a	yes
Staartjes et al [97]	2019	orthopedics	422	MCID	yes	yes	n/a	balanced	yes
Huber et al [98]	2019	orthopedics	34,110	yes	yes	no	n/a	partially balanced	yes
Wang et al [88]	2021	oncology	823	yes	no	no	no classification	n/a	yes
Xu et al [99]	2022	oncology	630	no	no	no	balanced	n/a	yes
Verna et al [147]	2022	orthopedics	1040	combined	yes	no	n/a	imbalanced	no
Harris et al [100]	2020	orthopedics	587	MCID	no	no	not reported	n/a	no
Siccoli et al [101]	2019	orthopedics	635	MCID	yes	yes	n/a	partially balanced	yes
Pfob et al [102]	2021	oncology	1553	combined	yes	no	n/a	imbalanced	yes
Curtis et al [136]	2022	orthopedics	494	no	no	no	imbalanced	n/a	yes
Iivanainen et al [137]	2021	oncology	34	yes	yes	no	n/a	imbalanced	no
Pedersen et al [103]	2022	orthopedics	1968	MCID	yes	yes	n/a	balanced	yes
Katakam et al [75]	2022	orthopedics	744	MCID	no	no	imbalanced	n/a	yes
Crowson et al [148]	2020	hearing	1604	yes	no	no	imbalanced	n/a	yes
Josephson et al [104]	2021	neurology	201	yes	no	no	imbalanced	n/a	yes
Ramkumar et al [105]	2021	orthopedics	153	MCID	no	no	imbalanced	n/a	no
Noel et al [138]	2022	oncology	11761	yes	yes	no	n/a	imbalanced	yes
Munn et al [106]	2022	orthopedics	1432	yes	no	no	imbalanced	n/a	yes
Kober et al [107]	2021	oncology	1217	yes	no	no	not reported	n/a	yes
Durand et al [108]	2020	orthopedics	1503	no	no	no	imbalanced	n/a	yes
Harrison et al [109]	2022	orthopedics	1916	yes	no	no	balanced	n/a	yes
Kumar et al [110]	2020	orthopedics	2887	yes	no	no	balanced	n/a	yes
Wshah et al [111]	2019	mental health	90	yes	no	no	not reported	n/a	yes
Lu et al [112]	2022	orthopedics	381	yes	no	no	imbalanced	n/a	yes
Kunze et al [124]	2021	orthopedics	818	MCID	no	no	imbalanced	n/a	yes
Kalweit et al [89]	2021	orthopedics	9500	yes	no	no	not reported	n/a	yes
Milella et al [149]	2022	orthopedics	6521	MCID	yes	yes	n/a	partially balanced	yes
Loos et al [114]	2022	orthopedics	2224	MCID	yes	yes	n/a	balanced	no
Hsu et al [167]	2022	neurology	3495	no	no	no	imbalanced	n/a	no
Iivanainen et al [160]	2022	oncology	31	no	yes	no	n/a	imbalanced	no
Polce et al [150]	2021	orthopedics	331	yes	no	no	imbalanced	n/a	yes
Chmiel et al [139]	2022	respiratory	2374	yes	no	no	imbalanced	n/a	yes
Sun et al [163]	2021	orthopedics	1,198	no	yes	yes	n/a	partially balanced	yes
Xuyi et al [278]	2021	oncology	46,104	yes	no	no	not reported	n/a	yes
Shipton-Sharman et al [115]	2022	neurology	2581	yes	no	no	imbalanced	n/a	no
Long et al [169]	2017	orthopedics	88	no	no	no	imbalanced	n/a	no
Luo et al [161]	2015	respiratory	210	yes	yes	yes	n/a	partially balanced	no
Ziobrowski et al [140]	2021	mental health	1003	yes	no	no	imbalanced	n/a	yes
Tan et al [116]	2013	respiratory	72	no	no	no	balanced	n/a	yes
Miranda et al [168]	2021	respiratory	217,580	no	yes	yes	n/a	partially balanced	no
Annapureddy et al [141]	2020	mental health	83	yes	yes	yes	n/a	partially balanced	yes
Thanathamthee [151]	2014	mental health	3115	yes	no	no	imbalanced	n/a	no
Mendoza et al [152]	2022	respiratory	1434868	no	yes	yes	n/a	balanced	no
Rahman et al [117]	2022	mental health	775	yes	no	no	not reported	n/a	no
Wang et al [265]	2010	sub-health state	572	no	yes	no	n/a	imbalanced	yes
Lee [118]	2017	mental health	9,089	no	no	no	imbalanced	n/a	yes
Xu et al [153]	2022	mental health	25214	yes	no	no	not reported	n/a	no
Zhong et al [142]	2022	mental health	915	yes	yes	no	n/a	imbalanced	yes
Pfob et al [279]	2022	oncology	1921	yes	no	no	imbalanced	n/a	yes
Van der Stap [173]	2022	oncology	532	yes	no	no	no classification	n/a	yes
Sidey-Gibbons et al [19]	2022	oncology	245	no	yes	yes	n/a	partially balanced	yes
Nowinka et al [119]	2022	orthopedics	5947	yes	yes	yes	n/a	imbalanced	yes
Arkin et al [120]	2020	oncology	189	no	no	no	imbalanced	n/a	no
Fu et al [154]	2018	oncology	355	no	no	no	balanced	n/a	no
Schultebrucks et al [164]	2021	mental health	473	yes	yes	yes	n/a	balanced	yes
Hoogendam et al [121]	2022	orthopedics	2119	MCID	yes	yes	n/a	balanced	yes
Camp et al [122]	2022	neurology	453	MCID	yes	no	n/a	imbalanced	yes
Canas et al [143]	2021	respiratory	182991	no	yes	no	n/a	imbalanced	no
Wardenaar [123]	2021	mental health	2,981	yes	no	no	imbalanced	n/a	yes
Bugajski [172]	2021	respiratory	20	yes	no	no	not reported	n/a	no
Sim et al [6]	2020	oncology	809	no	yes	yes	n/a	balanced	yes
Pan et al [155]	2020	oncology	86	yes	no	no	imbalanced	n/a	yes
Kunze et al [113]	2020	orthopedics	616	MCID	no	no	imbalanced	n/a	yes
Sutradhar [144]	2020	oncology	212615	no	no	no	not reported	n/a	no
Verna et al [280]	2022	orthopedics	293	yes	no	no	no classification	n/a	no
Smith et al [125]	2022	mental health	361	yes	no	no	no classification	n/a	yes
Agochukwu-Mmonu et al [156]	2022	oncology	2653	yes	no	no	imbalanced	n/a	yes
Goldstein et al [145]	2021	respiratory	116	no	yes	yes	n/a	balanced	no
Bone et al [126]	2021	mental health	2317	yes	no	no	balanced	n/a	yes
O'Driscoll et al [127]	2021	mental health	2858	yes	no	no	not reported	n/a	yes
Tschuggnall et al [76]	2021	orthopedics	1047	combined	no	no	not reported	n/a	yes
Rajpurkar et al [34]	2020	mental health	518	yes	no	no	not reported	n/a	no
Shahzad et al [171]	2020	mental health	600	yes	no	no	imbalanced	n/a	no
Shi et al [128]	2012	oncology	402	yes	no	no	imbalanced	n/a	no
Pua et al [129]	2020	orthopedics	4026	yes	no	no	imbalanced	n/a	yes
Chekroud et al [157]	2016	mental health	4041	yes	no	no	not reported	n/a	yes
Nijeweme-D'Hollosy et al [158]	2018	orthopedics	1288	no	no	no	imbalanced	n/a	no
Rahman et al [146]	2019	orthopedics	31,700	yes	yes	yes	n/a	balanced	no
Langenberger et al [135]	2022	orthopedics	1,843	MCID	yes	no	n/a	n/a	yes
Pappot et al [166]	2023	oncology	238	no	no	no	imbalanced	n/a	yes
Hasannejadsl et al [131]	2023	oncology	964	yes	yes	yes	n/a	partially balanced	yes
Xu et al [132]	2023	oncology	3,058	yes	yes	no	n/a	n/a	yes
Grant et al [90]	2023	oncology	105,129	no	no	no	balanced	n/a	yes
Tian et al [133]	2023	cardiovascular	941	no	yes	yes	n/a	partially balanced	yes
Park et al [134]	2023	orthopedics	1,141	MCID	no	no	balanced	n/a	yes
Goldstein et al [159]	2023	endometriosis	886	no	no	no	balanced	n/a	no
Langenberger [130]	2023	orthopedics	5,934	no	no	no	balanced	n/a	yes

Table A3: Model development and evaluation, including study characteristics of papers included in the review (Chapter 2).

Authors	Hyperparameter tuning	Hyperparameters reported?	Variable importance reported?	Best performing model	Clinicians involvement?	Country	Stage of development	Age reported?	Gender reported?	Ethnicity reported?
Klent et al [91]		no	yes	NN	no	USA	internal validation	yes	yes	no
Saadhani et al [165]	experimentally selected	no	yes	NN	no	New Zealand	internal validation	no	yes	yes
Tennenhouse et al [92]	done, no technique	no	yes	regression	yes	Canada	internal validation	yes	yes	yes
Zhang et al [93]	cross-validation	no	yes	regression	no	Singapore	internal validation	yes	yes	no
Zhang et al [94]	cross-validation	no	yes	regression	no	Singapore	internal validation	yes	yes	yes
Haeblerle et al [162]	nr	no	yes	n/a	no	USA	internal validation	yes	yes	yes
Verna et al [95]	grid search	no	yes	regression	no	Norway	internal validation	no	no	no
Buis et al [96]	grid search	no	no	RF	no	Denmark	internal validation	yes	yes	no
Staarjies et al [97]	grid search	no	yes	NN	no	Netherlands	internal validation	yes	yes	no
Huber et al [98]	grid search	no	yes	boosting	no	Germany	internal validation	yes	yes	no
Wang et al [88]	nr	no	no	LSTM	no	USA	internal validation	no	no	no
Xu et al [99]	Bayesian Optimisation	yes	yes	boosting	no	USA	internal validation	yes	yes	yes
Verna et al [147]	random search	yes	yes	regression	no	Netherlands	internal validation	no	no	no
Harris et al [100]	nr	no	no	QDA	no	USA	external validation	yes	yes	yes
Siccoli et al [101]	done, no technique	no	yes	nr	no	Netherlands	internal validation	yes	yes	no
Pfob et al [102]	grid search	yes	yes	NN	no	USA	internal validation	yes	no	yes
Curtis et al [136]	random search	no	yes	RF	no	USA and Canada	internal validation	yes	yes	yes
Ivanainen et al [137]	grid search	no	yes	boosting	no	Finland	internal validation	no	no	no
Pedersen et al [103]	done, no technique	no	no	SVM	no	Danemark	internal validation	yes	yes	no
Katakam et al [75]	nr	no	yes	regression	no	USA	internal validation	yes	yes	no
Crowson et al [148]	nr	no	yes	NN	no	Canada	internal validation	yes	yes	no
Josephson et al [104]	nr	no	no	nc	no	Canada and Sweden	internal validation	yes	yes	no
Ramkumar et al [105]	nr	no	yes	NN	no	USA	internal validation	yes	yes	yes
Noel et al [138]	grid search	no	yes	boosting	no	USA	internal validation	yes	yes	no
Munn et al [106]	nr	no	yes	regression	no	Canada	internal validation	yes	yes	no
Kober et al [107]	nr	no	yes	RF	no	USA	internal validation	yes	yes	yes
Durand et al [108]	grid search	no	yes	SVM	no	USA	internal validation	yes	yes	no
Harrison et al [109]	nr	no	no	regression	no	Scotland	internal validation	yes	yes	no
Kumar et al [110]	nr	no	yes	Wide and deep	no	USA	internal validation	yes	yes	no
Wahab et al [111]	grid search	yes	yes	SVM	no	USA	internal validation	yes	yes	yes
Lu et al [112]	nr	no	yes	RF	no	USA	internal validation	yes	yes	no
Kunze et al [124]	nr	no	yes	boosting	no	USA	internal validation	yes	yes	yes
Kalweit et al [89]	random search	yes	yes	AdaptiveNet	no	Switzerland	internal validation	yes	no	no
Milella et al [149]	random search	yes	no	boosting	no	Italy	internal validation	yes	yes	no
Loos et al [114]	nr	no	no	boosting	no	Netherlands	internal validation	yes	yes	no
Hsu et al [167]	nr	no	yes	SVM	yes	Taiwan	internal validation	yes	yes	no
Ivanainen et al [160]	nr	no	yes	boosting	no	Finland	internal validation	yes	yes	yes
Polev et al [150]	nr	no	yes	SVM	no	USA	internal validation	yes	yes	no
Chmiel et al [139]	Bayesian Optimisation	no	yes	RF	no	UK	internal validation	yes	yes	no
Sun et al [163]	nr	yes	yes	n/a	no	USA	internal validation	yes	yes	yes
Xuyi et al [278]	nr	yes	no	n/a	no	Canada	internal validation	yes	yes	no
Shipton-Sharma et al [115]	experimentally selected	yes	no	regression	no	Scotland	internal validation	yes	yes	no
Long et al [169]	cross-validation	no	yes	nc	no	UK	internal validation	yes	yes	no
Luo et al [161]	done, no technique	no	no	boosting	no	USA	internal validation	yes	yes	yes
Zolotorowski et al [140]	super learner	no	yes	nc	no	USA	internal validation	yes	yes	yes
Tan et al [116]	nr	no	yes	n/a	no	USA	internal validation	no	no	no
Miranda et al [168]	cross-validation	yes	no	NN	no	Brazil	internal validation	yes	yes	yes
Annappureddy et al [141]	Bayesian Optimisation	no	no	voting classifier	no	USA	internal validation	no	yes	no
Thanathamthee [151]	nr	no	yes	boosting	no	Thailand	internal validation	no	no	no
Mendoza et al [152]	nr	no	yes	boosting	no	Philippines	internal validation	no	no	no
Rahman et al [17]	random search	no	yes	regression	no	Jordan	internal validation	no	no	no
Wang et al [205]	nr	no	yes	n/a	no	China	internal validation	no	no	no
Lee [118]	nr	no	yes	regression	no	Korea	internal validation	yes	yes	no
Xu et al [153]	nr	yes	yes	voting classifier	no	China	internal validation	no	no	no
Zhong et al [142]	grid search	yes	yes	boosting	no	Sweden	internal validation	no	no	no
Pfob et al [279]	nr	no	yes	regression	no	USA and Canada	internal validation	yes	yes	yes
Van der Stap [173]	nr	no	no	n/a	no	Netherlands	internal validation	yes	yes	no
Sider-Gibbons et al [19]	grid search	yes	yes	NN	no	USA	internal validation	yes	yes	yes
Nowinski et al [119]	done, no technique	no	yes	regression	no	USA	external validation	yes	yes	yes
Arkin et al [120]	nr	no	yes	NN	no	Turkey	internal validation	yes	yes	no
Fu et al [154]	cross-validation	yes	no	NN	no	USA	internal validation	yes	yes	yes
Schultebras et al [164]	random search	yes	yes	SVM	no	USA	internal validation	yes?	?	no
Hoogendam et al [121]	cross-validation	no	yes	regression	no	Netherlands	deployment	yes	yes	no
Camp et al [122]	grid search	no	yes	SVM	no	USA	deployment	yes	yes	yes
Canas et al [143]	nr	no	yes	hierarchical gaussian process	no	UK and USA	internal validation	yes	yes	no
Wardman [123]	nr	no	yes	SuperLearner	no	Netherlands	internal validation	yes	yes	yes
Bugajski [172]	nr	yes	no	n/a	no	USA	internal validation	yes	yes	yes
Sim et al [6]	done, no technique	no	yes	bagging	no	South Korea	internal validation	yes?	?	?
Pan et al [155]	nr	no	no	boosting	no	USA	external validation	no	no	no
Kunze et al [113]	nr	no	yes	RF	no	USA	deployment	yes	no	no
Sutradhar [144]	nr	yes	no	NN	no	Canada	internal validation	yes	yes	no
Verna et al [280]	grid search	no	yes	boosting	no	Norway	external validation	no	no	no
Smith et al [125]	nr	no	no	regression	no	USA	external validation	yes	yes	yes
Agochukwu-Mmonim et al [156]	nr	no	yes	n/a	no	USA	internal validation	yes	yes	no
Goldstein et al [145]	nr	no	yes	RF	no	Israel	internal validation	yes	yes	no
Bone et al [126]	nr	yes	no	boosting	no	UK	external validation	yes	yes	no
O Driscoll et al [127]	nr	yes	yes	regression	no	UK	internal validation	yes	yes	yes
Tschuggnall et al [76]	cross-validation	no	yes	RF	no	Austria	internal validation	no	no	no
Rajpurkar et al [34]	grid search	no	yes	n/a	no	UK and USA	internal validation	yes	yes	no
Shahzad et al [171]	experimentally selected	no	yes	n/a	no	Pakistan	internal validation	no	yes	no
Shi et al [128]	nr	yes	yes	NN	no	Taiwan	internal validation	yes	no	no
Pua et al [129]	cross-validation	no	yes	regression	no	Singapore	internal validation	yes	yes	yes
Chekroun et al [157]	cross-validation	no	yes	nc	no	USA	external validation	no	no	no
Nijewene-D'Holloy et al [158]	nr	no	yes	boosting	yes	Netherlands	internal validation	yes	no	no
Rahman et al [146]	nr	no	yes	regression	no	UK	internal validation	no	no	no
Langenberger et al [135]	cross-validation	yes	yes	RF, regression, boosting	no	Germany	internal validation	yes	yes	no
Pappot et al [166]	nr	yes	yes	random forest	no	Denmark	internal validation	no	yes	no
Hasanajejad et al [131]	nr	no	yes	LR	no	Netherlands	external validation	yes	yes	no
Xu et al [132]	hypergrid search	yes	yes	LR	no	USA	external validation	no	no	no
Grant et al [90]	Bayesian Optimisation	no	yes	voting classifier	no	Canada	external validation	yes	yes	no
Tian et al [133]	search	yes	yes	XGBoost	no	China	internal validation	yes	yes	no
Park et al [134]	nr	yes	yes	regression	no	USA	internal validation	yes	yes	yes
Goldstein et al [159]	nr	no	yes	AdaBoost	no	Europe, US, Australia, Israel	internal validation	no	no	no
Langenberger [130]	grid search	no	yes	Xboost and regression	no	Germany	internal validation	yes	yes	no

Section/Topic	Item	Development / evaluation ¹	Checklist item	Reported on page
TITLE				
<i>Title</i>	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	Chapter titles
ABSTRACT				
<i>Abstract</i>	2	D;E	See TRIPOD+AI for Abstracts checklist	
INTRODUCTION				
<i>Background</i>	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	Chapter 1, section 1.5
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	Chapter 3, section 3.3
	3c	D;E	Describe any known health inequalities between sociodemographic groups	Chapter 8, Section 8.6
<i>Objectives</i>	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	Chapter 1, section 1.5
METHODS				
<i>Data</i>	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	Chapter 3, section 3.4
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	Chapter 3, section 3.3 and 3.4
<i>Participants</i>	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	Chapter 3, section 3.3
	6b	D;E	Describe the eligibility criteria for study participants	Referenced
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	Chapter 3, section 3.3
<i>Data preparation</i>	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	Methods sections, chapters 4.5,6,7
<i>Outcome</i>	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	Methods sections, chapters 4.5,6,7
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	Chapter 3, section 3.3
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	N/A
<i>Predictors</i>	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	Chapter 3, section 3.3
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	Methods sections, chapters 4.5,6,7
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	Chapter 4, Table 4.2
<i>Sample size</i>	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	Methods sections, chapters 4.5,6,7
<i>Missing data</i>	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	Methods sections, chapters 4.5,6,7
<i>Analytical methods</i>	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	Methods sections, chapters 4.5,6,7
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	Methods sections, chapters 4.5,6,7
	12c	D	Specify the type of model, rationale ² , all model-building steps, including any hyperparameter tuning, and method for internal validation	Methods sections, chapters 4.5,6,7
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations ³	N/A
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	Methods sections, chapters 4.5,6,7
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	Methods sections, chapters 4.5,6,7
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	Methods sections, chapters 4.5,6,7
<i>Class imbalance</i>	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	Methods sections, chapters 4.5,6,7
<i>Fairness</i>	14	D;E	Describe any approaches that were used to address model fairness and their rationale	Methods sections, chapters 6,7
<i>Model output</i>	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	Methods sections, chapters 4.5,6,7

¹ D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model

² Separately for all model building approaches.

³ TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Figure A3: TRIPOD+AI checklist part 1 (Chapter 3).

<i>Training versus evaluation</i>	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	Methods sections, chapters 4,5,6,7
<i>Ethical approval</i>	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	Chapter 3, section 3.3
OPEN SCIENCE				
<i>Funding</i>	18a	D;E	Give the source of funding and the role of the funders for the present study	Acknowledgements
<i>Conflicts of interest</i>	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	N/A
<i>Protocol</i>	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	Chapter 2, methods section
<i>Registration</i>	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	N/A
<i>Data sharing</i>	18e	D;E	Provide details of the availability of the study data	N/A (ethics)
<i>Code sharing</i>	18f	D;E	Provide details of the availability of the analytical code ⁴	N/A (ethics)
PATIENT & PUBLIC INVOLVEMENT				
<i>Patient & Public Involvement</i>	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	Methods sections, chapters 4,5,6,7
RESULTS				
<i>Participants</i>	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Chapter 4, Table 4.2
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	Methods sections, chapters 4,5,6,7
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	Results sections, chapters 4,5,6,7
<i>Model development</i>	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	Methods sections, chapters 4,5,6,7
<i>Model specification</i>	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary) ⁵	Methods sections, chapters 4,5,6,7
<i>Model performance</i>	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	Results sections, chapters 4,5,6,7
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details ³ .	N/A
<i>Model updating</i>	24	E	Report the results from any model updating, including the updated model and subsequent performance	N/A
DISCUSSION				
<i>Interpretation</i>	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	Discussion sections, chapters 4,5,6,7
<i>Limitations</i>	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	Discussion sections, chapters 4,5,6,7
<i>Usability of the model in the context of current care</i>	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	Discussion sections, chapters 4,5,6,7
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	Chapter 8
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	Chapter 8

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

⁴ This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.

⁵ This relates to the code to implement the model to get estimates of risk for a new individual.

Figure A4: TRIPOD+AI checklist part 2 (Chapter 3).

		LOGISTIC REGRESSION																	
		original data						balanced data											
		accuracy	precision	recall	F1	AUC	hyperparameters	accuracy	precision	recall	F1	AUC	hyperparameters						
admissions	clinical variables	0.647	0.627	0.654	0.61	0.513	c=10, solver=lbfgs	0.654	0.7	0.653	0.676	0.555	c=10, solver=lbfgs						
	all variables	0.738	0.476	0.476	0.476	0.605	c=100, solver=lbfgs	0.694	0.721	0.721	0.721	0.681	c=100, solver=lbfgs						
	PRIMA	0.75	0.5	0.238	0.233	0.579	c=100, solver=lbfgs	0.613	0.607	0.59	0.626	0.616	c=1, solver=lbfgs						
	clinical + CBO	0.75	0.5	0.476	0.488	0.659	c=10, solver=lbfgs	0.677	0.727	0.658	0.69	0.678	c=100, solver=lbfgs						
	clinical + EQ-SD	0.637	0	0	0	0.5	c=100, solver=lbfgs	0.64	0.703	0.6	0.647	0.644	c=100, solver=lbfgs						
triage	clinical + FACT-6	0.647	0.505	0.502	0.502	0.543	c=1, solver=lbfgs	0.632	0.607	0.613	0.616	0.635	c=1, solver=lbfgs						
	clinical variables	0.588	0.641	0.783	0.784	0.522	c=100, solver=lbfgs	0.568	0.571	0.6	0.585	0.587	c=100, solver=lbfgs						
	all variables	0.631	0.655	0.824	0.73	0.578	c=10, solver=lbfgs	0.573	0.491	0.65	0.559	0.584	c=1, solver=lbfgs						
	PRIMA	0.643	0.667	0.824	0.797	0.594	c=10, solver=lbfgs	0.594	0.514	0.475	0.494	0.577	c=1, solver=lbfgs						
	clinical + CBO	0.631	0.667	0.794	0.721	0.589	c=10, solver=lbfgs	0.594	0.51	0.625	0.582	0.598	c=1, solver=lbfgs						
chemo	clinical + EQ-SD	0.578	0.644	0.734	0.686	0.525	c=10, solver=lbfgs	0.558	0.515	0.583	0.547	0.567	c=1, solver=lbfgs						
	clinical + FACT-6	0.598	0.646	0.787	0.723	0.513	c=1, solver=lbfgs	0.565	0.527	0.689	0.626	0.643	c=1, solver=lbfgs						
	clinical variables	0.647	0.712	0.775	0.743	0.588	c=10, solver=lbfgs	0.642	0.565	0.684	0.619	0.647	c=1, solver=lbfgs						
	all variables	0.69	0.75	0.778	0.764	0.656	c=10, solver=lbfgs	0.732	0.7	0.778	0.737	0.734	c=1, solver=lbfgs						
	PRIMA	0.653	0.681	0.87	0.794	0.589	c=1, solver=lbfgs	0.58	0.566	0.556	0.561	0.58	c=100, solver=lbfgs						
	clinical + CBO	0.714	0.742	0.852	0.799	0.659	c=10, solver=lbfgs	0.688	0.666	0.741	0.696	0.689	c=1, solver=lbfgs						
	clinical + EQ-SD	0.647	0.653	0.985	0.786	0.493	c=100, solver=lbfgs	0.675	0.594	0.719	0.651	0.678	c=1, solver=lbfgs						
	clinical + FACT-6	0.608	0.69	0.731	0.71	0.551	c=1, solver=lbfgs	0.627	0.552	0.689	0.597	0.613	c=1, solver=lbfgs						
		DECISION TREE																	
		original data						balanced data											
accuracy	precision	recall	F1	AUC	hyperparameters	accuracy	precision	recall	F1	AUC	hyperparameters	accuracy	precision	recall	F1	AUC	hyperparameters		
admissions	clinical variables	0.618	0.25	0.027	0.049	0.49	gini, 2, 2	0.75	0.773	0.773	0.747	0.747	gini, 16, 2, 4	0.598	0.308	0.25	0.267	0.485	gini, 16, 2, 4
	all variables	0.738	0.444	0.19	0.267	0.558	entropy, 16, 4, 2	0.759	0.779	0.889	0.842	0.749	gini, 16, 12, 2, 4	0.738	0.444	0.19	0.267	0.558	gini, 16, 12, 2, 4
	PRIMA	0.738	0.444	0.19	0.267	0.558	gini, 2, 2, 2	0.689	0.489	0.398	0.45	0.646	gini, 16, 12, 2, 4	0.689	0.489	0.398	0.45	0.646	gini, 16, 12, 2, 4
	clinical + CBO	0.631	0.521	0.429	0.567	0.563	entropy, 8, 16, 4	0.62	0.66	0.853	0.821	0.822	gini, 16, 2, 2	0.659	0.433	0.464	0.488	0.558	gini, 16, 2, 2
	clinical + EQ-SD	0.618	0.644	0.552	0.61	0.556	gini, 2, 2, 2	0.757	0.733	0.69	0.71	0.749	gini, 16, 12, 2, 4	0.757	0.733	0.69	0.71	0.749	gini, 16, 12, 2, 4
triage	clinical + FACT-6	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.779	0.785	0.847	0.865	0.774	gini, 14, 2, 2	0.578	0.32	0.235	0.271	0.493	gini, 14, 2, 2
	clinical variables	0.549	0.613	0.75	0.676	0.48	gini, 2, 2, 2	0.61	0.622	0.5	0.566	0.622	gini, 16, 16, 2, 4	0.51	0.576	0.756	0.756	0.497	gini, 16, 16, 2, 4
	all variables	0.738	0.444	0.19	0.267	0.558	gini, 16, 12, 2, 4	0.64	0.555	0.525	0.525	0.585	gini, 16, 12, 2, 4	0.571	0.485	0.542	0.591	0.576	gini, 16, 12, 2, 4
	PRIMA	0.607	0.686	0.647	0.657	0.596	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical + CBO	0.488	0.611	0.431	0.556	0.504	gini, 16, 12, 12, 15	0.647	0.615	0.6	0.606	0.666	gini, 16, 12, 2, 4	0.54	0.574	0.666	0.688	0.593	gini, 16, 12, 2, 4
chemo	clinical + EQ-SD	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical + FACT-6	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical variables	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	all variables	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	PRIMA	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical + CBO	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical + EQ-SD	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
	clinical + FACT-6	0.598	0.612	0.578	0.594	0.535	entropy, 16, 4, 2	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4	0.651	0.655	0.655	0.655	0.655	gini, 16, 12, 2, 4
		RANDOM FOREST																	
		original data						balanced data											
accuracy	precision	recall	F1	AUC	hyperparameters	accuracy	precision	recall	F1	AUC	hyperparameters	accuracy	precision	recall	F1	AUC	hyperparameters		
admissions	clinical variables	0.618	0.25	0.027	0.049	0.49	SAMME, 0.1, 100-500	0.646	0.629	0.52	0.589	0.571	SAMME, 1, 0, 100-500	0.646	0.646	0.536	0.5	0.616	SAMME, 1, 0, 100-500
	all variables	0.738	0.333	0.248	0.318	0.528	SAMME, 0.1, 100-500	0.649	0.683	0.572	0.578	0.549	SAMME, 1, 0, 100-500	0.548	0.353	0.249	0.287	0.512	SAMME, 1, 0, 100-500
	PRIMA	0.738	0.333	0.248	0.318	0.528	SAMME, 0.1, 100-500	0.649	0.683	0.572	0.578	0.549	SAMME, 1, 0, 100-500	0.548	0.353	0.249	0.287	0.512	SAMME, 1, 0, 100-500
	clinical + CBO	0.617	0.58	0.293	0.332	0.546	SAMME, 0.1, 100-500	0.739	0.737	0.87	0.797	0.739	SAMME, 0.1, 100-500	0.595	0.417	0.336	0.348	0.548	SAMME, 1, 0, 100-500
	clinical + EQ-SD	0.617	0.58	0.293	0.332	0.546	SAMME, 0.1, 100-500	0.61	0.63	0.547	0.607	0.638	SAMME, 1, 0, 100-500	0.461	0.238	0.235	0.225	0.404	SAMME, 1, 0, 100-500
triage	clinical + FACT-6	0.628	0.25	0.027	0.049	0.49	SAMME, 0.1, 100-500	0.644	0.716	0.707	0.711	0.683	SAMME, 1, 0, 100-500	0.48	0.229	0.235	0.232	0.419	SAMME, 1, 0, 100-500
	clinical variables	0.549	0.613	0.75	0.676	0.48	SAMME, 0.1, 100-500	0.66	0.585	0.512	0.529	0.589	SAMME, 1, 0, 100-500	0.51	0.579	0.759	0.759	0.497	SAMME, 1, 0, 100-500
	all variables	0.512	0.586	0.647	0.624	0.47	SAMME, 0.1, 100-500	0.667	0.635	0.7	0.636	0.647	SAMME, 1, 0, 100-500	0.476	0.544	0.521	0.523	0.469	SAMME, 1, 0, 100-500
	PRIMA	0.524	0.604	0.627	0.615	0.496	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	clinical + CBO	0.497	0.591	0.792	0.695	0.509	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
chemo	clinical + EQ-SD	0.578	0.643	0.797	0.703	0.504	SAMME, 0.01, 100-500	0.528	0.536	0.5	0.517	0.526	SAMME, 1, 0, 100-500	0.578	0.648	0.732	0.661	0.554	SAMME, 0.1, 100-500
	clinical + FACT-6	0.578	0.643	0.797	0.703	0.504	SAMME, 0.01, 100-500	0.528	0.536	0.5	0.517	0.526	SAMME, 1, 0, 100-500	0.578	0.648	0.732	0.661	0.554	SAMME, 0.1, 100-500
	clinical variables	0.549	0.613	0.75	0.676	0.48	SAMME, 0.1, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	all variables	0.512	0.586	0.647	0.624	0.47	SAMME, 0.1, 100-500	0.667	0.635	0.7	0.636	0.647	SAMME, 1, 0, 100-500	0.476	0.544	0.521	0.523	0.469	SAMME, 1, 0, 100-500
	PRIMA	0.524	0.604	0.627	0.615	0.496	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	clinical + CBO	0.497	0.591	0.792	0.695	0.509	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	clinical + EQ-SD	0.578	0.643	0.797	0.703	0.504	SAMME, 0.1, 100-500	0.528	0.536	0.5	0.517	0.526	SAMME, 1, 0, 100-500	0.578	0.648	0.732	0.661	0.554	SAMME, 0.1, 100-500
	clinical + FACT-6	0.578	0.643	0.797	0.703	0.504	SAMME, 0.1, 100-500	0.528	0.536	0.5	0.517	0.526	SAMME, 1, 0, 100-500	0.578	0.648	0.732	0.661	0.554	SAMME, 0.1, 100-500
	clinical variables	0.549	0.613	0.75	0.676	0.48	SAMME, 0.1, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	all variables	0.512	0.586	0.647	0.624	0.47	SAMME, 0.1, 100-500	0.667	0.635	0.7	0.636	0.647	SAMME, 1, 0, 100-500	0.476	0.544	0.521	0.523	0.469	SAMME, 1, 0, 100-500
	PRIMA	0.524	0.604	0.627	0.615	0.496	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	clinical + CBO	0.497	0.591	0.792	0.695	0.509	SAMME, 1, 0, 100-500	0.646	0.56	0.7	0.622	0.634	SAMME, 1, 0, 100-500	0.464	0.532	0.521	0.526	0.465	SAMME, 1, 0, 100-500
	clinical + EQ-SD	0.578	0.643	0.797	0.703	0.504	SAMME, 0.1, 100-500	0.528	0.536	0.5	0.517	0.526	SAMME, 1, 0, 100-500	0.578	0.648	0.732	0.661	0.554	SAMME, 0.1, 100-500

Table A4: Hyperparameters of each model predicting deterioration of the HRQoL outcomes (Chapter 5).

Outcome	Outcome at	Model	Hyperparameters
Physical well-being	6 weeks	LR	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 100
		MLP	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Social functioning	6 weeks	LR	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'sgd'
Outcome	Outcome at	Model	Hyperparameters
Role functioning	6 weeks	LR	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion: 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'sgd'
	12 weeks	LR	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C: 0.1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 100
		MLP	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100,), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Usual activities	6 weeks	LR	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'tanh', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100,), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C: 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation: 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Mobility	6 weeks	LR	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C: 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion: 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate: 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C: 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'lbfgs'
		RF	criterion: 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate: 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation: 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'

Table A5: Hyperparameters of each model predicting improvement of the HRQoL outcomes (Chapter 5).

Outcome	Outcome at	Model	Hyperparameters
Physical well-being	6 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion': 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 100
		MLP	activation': 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'lbfgs'
		RF	criterion': 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'lbfgs'
		RF	criterion': 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 0.1, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Social functioning	6 weeks	LR	C': 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion': 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation': 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'entropy', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (100,), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 500
		XGB	learning_rate': 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Role functioning	6 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion': 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 0.001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation': 'tanh', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C': 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'tanh', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Usual activities	6 weeks	LR	C': 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion': 'gini', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C': 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'newton-cg'
		RF	criterion': 'gini', 'max_depth': 14, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 100
		MLP	activation': 'relu', 'alpha': 0.01, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
Outcome	Outcome at	Model	Hyperparameters
Mobility	6 weeks	LR	C': 1, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'entropy', 'max_depth': 18, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (100, 50, 20), 'max_iter': 10000, 'solver': 'adam'
	12 weeks	LR	C': 100, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'gini', 'max_depth': 8, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 1.0, 'loss': 'log_loss', 'n_estimators': 100
		MLP	activation': 'relu', 'alpha': 1e-05, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'
	18 weeks	LR	C': 10, 'max_iter': 10000, 'penalty': 'l2', 'solver': 'liblinear'
		RF	criterion': 'gini', 'max_depth': 8, 'min_samples_split': 2, 'n_estimators': 100
		XGB	learning_rate': 0.1, 'loss': 'exponential', 'n_estimators': 500
		MLP	activation': 'relu', 'alpha': 0.0001, 'early_stopping': True, 'hidden_layer_sizes': (150, 60, 30), 'max_iter': 10000, 'solver': 'adam'

Table A6: Feature importance ranks for physical well-being changes at 18 weeks prediction models LR and RF (Chapter 5).

Physical well-being							
Deterioration				Improvement			
Logistic regression		Random forest		Logistic regression		Random forest	
PWB_overall0	1.569433	PWB_overall0	0.133532	CMResCOPD	13.923359	PWB_overall0	0.221759
CMRheuArth	1.149526	BCBMI	0.09927	CMCarMI	13.066001	BCBMI	0.073412
CMStomInflamm	1.026601	FWB_overall0	0.081032	CMCarAngina	9.672283	EQ5DPain0	0.071416
PrimaryorMet	0.879178	SWB_overall0	0.075772	CMEndDiabetes	4.327627	EQ5DVAS0	0.064324
PreviousChemo	0.864974	EQ5DVAS0	0.066705	CMEndHyperth	4.285474	SWB_overall0	0.062971
CMPrevMal	0.79102	DiseaseSite	0.050577	CMNeuStroke	3.695441	FWB_overall0	0.062633
CMResCOPD	0.764014	DCEmployment	0.050358	CMEndHypo	3.601606	ROLO	0.050095
CMCarMI	0.59501	ROLO	0.050047	Comorbidities	2.922004	DiseaseSite	0.045489
DiseaseSite	0.564434	PrimaryorMet	0.041654	CMCarHyperten	2.919881	ed_lev	0.038387
CMEndDiabetes	0.501928	ed_lev	0.039405	PWB_overall0	2.850354	PrimaryorMet	0.035688
CMRheuCTD	0.444377	EQ5DAnxDep0	0.034651	CMCarVenous	2.747976	DCEmployment	0.03415
FWB_overall0	0.444371	DCMarital	0.030691	CMCarArrhythm	2.534306	DCMarital	0.028505
CMEndHypo	0.426059	EQ5DPain0	0.028751	PreviousChemo	2.414958	EQ5DUUsAct0	0.026227
EQ5DAnxDep0	0.383677	EQ5DUUsAct0	0.028591	CMRheuArth	1.93644	Comorbidities	0.025013
EQ5DPain0	0.364803	Comorbidities	0.028133	PrimaryorMet	1.54155	EQ5DAnxDep0	0.024492
CMCarVenous	0.332789	EQ5DMob0	0.021961	CMResAsthma	1.499618	EQ5DMob0	0.020482
CMResAsthma	0.289044	StudyArm	0.019965	CMPrevMal	1.319404	StudyArm	0.020222
StudyArm	0.28145	PreviousChemo	0.016012	CMSubstAlcohol	1.042471	PreviousChemo	0.013142
BCBMI	0.252698	CMCarHyperten	0.015805	DiseaseSite	0.893342	CMEndDiabetes	0.013136
CMEndHyperth	0.243022	CMResCOPD	0.012084	ed_lev	0.864195	CMCarHyperten	0.01048
CMGasPancreas	0.228956	CMResAsthma	0.011007	EQ5DVAS0	0.832183	CMEndHypo	0.009797
ROLO	0.224995	EQ5DSelCar0	0.010529	EQ5DSelCar0	0.796896	EQ5DSelCar0	0.009597
CMStomUlcers	0.200148	CMPrevMal	0.009927	CMGasPancreas	0.74042	CMPrevMal	0.008038
CMNeuStroke	0.197854	CMEndDiabetes	0.009848	CMRheuLupus	0.703934	CMResAsthma	0.006365
CMRenEndStage	0.178679	CMCarMI	0.006526	BCBMI	0.599315	CMCarAngina	0.004794
ed_lev	0.135084	CMEndHypo	0.006286	FWB_overall0	0.560101	CMCarMI	0.003373
Comorbidities	0.127383	CMStomInflamm	0.004816	CMNeuParkins	0.392506	CMEndHyperth	0.002924
CMCarAngina	0.123243	CMCarVenous	0.004777	CMStomInflamm	0.349838	CMCarVenous	0.002866
EQ5DVAS0	0.110928	CMCarAngina	0.002398	DCEmployment	0.189108	CMNeuParkins	0.00223
CMCarArrhythm	0.106706	CMNeuStroke	0.001957	DCMarital	0.180942	CMResCOPD	0.001576
SWB_overall0	0.102555	CMRheuArth	0.001828	EQ5DMob0	0.171249	CMCarArrhythm	0.001128
DCEmployment	0.100668	CMRheuLupus	0.001095	CMStomUlcers	0.149692	CMRheuArth	0.001108
EQ5DMob0	0.081411	CMCarArrhythm	0.000916	ROLO	0.100497	CMStomInflamm	0.001044
CMCarHeartFail	0.078057	CMRheuCTD	0.000801	SWB_overall0	0.099762	CMNeuStroke	0.000839
CMSubstAlcohol	0.076215	CMEndHyperth	0.000681	EQ5DAnxDep0	0.085732	CMGasPancreas	0.000631
DCMarital	0.07464	CMGasPancreas	0.000615	StudyArm	0.080389	CMSubstAlcohol	0.00056
EQ5DUUsAct0	0.022877	CMSubstAlcohol	0.000385	EQ5DUUsAct0	0.042851	CMRheuCTD	0.000482
CMCarHyperten	0.012993	CMRenEndStage	0.000346	EQ5DPain0	0.042741	CMRheuLupus	0.00029
EQ5DSelCar0	0.010388	CMStomUlcers	0.000227	CMRheuCTD	0.035837	CMStomUlcers	0.000284
CMRheuLupus	0.010197	CMCarHeartFail	0.00004	CMResChronBron	0.005954	CMResChronBron	0.000052
CMSubstDrugs	0	CMSubstDrugs	0	CMSubstDrugs	0	CMGasChronHep	0
CMResChronBron	0	CMGasChronHep	0	CMStomMalabsor	0	CMRheuRhPolymy	0
CMStomMalabsor	0	CMResChronBron	0	CMRheuRhPolymy	0	CMResEmphys	0
CMResEmphys	0	CMNeuMyasth	0	CMGasCirrrosis	0	CMRheuPolymyo	0
CMRheuRhPolymy	0	CMNeuParkins	0	CMRheuPolymyo	0	CMSubstDrugs	0
CMNeuMyasth	0	CMNeuMS	0	CMGasChronHep	0	CMNeuMyasth	0
CMGasChronHep	0	CMGasCirrrosis	0	CMResEmphys	0	CMNeuMS	0
CMNeuParkins	0	CMRheuRhPolymy	0	CMRenEndStage	0	CMCarHeartFail	0
CMGasCirrrosis	0	CMRheuPolymyo	0	CMCarHeartFail	0	CMGasCirrrosis	0
CMNeuMS	0	CMStomMalabsor	0	CMNeuMyasth	0	CMRenEndStage	0
CMRheuPolymyo	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor	0

Table A7: Feature importance ranks for usual activities changes at 18 weeks prediction models LR and RF (Chapter 5).

Usual activities							
Deterioration				Improvement			
Logistic regression		Random forest		Logistic regression		Random forest	
EQ5DUsuAct0	2.348261	BCBMI	0.143117	CMSubstAlcohol	7.569993	EQ5DUsuAct0	0.252763
CMRheuLupus	1.244129	EQ5DUsuAct0	0.129836	CMRheuArth	3.585515	ROL0	0.115769
DiseaseSite	0.898524	DiseaseSite	0.077861	EQ5DUsuAct0	3.522745	PWB_overall0	0.08537
CMPrevMal	0.865107	EQ5DVAS0	0.07398	CMCarVenous	2.355063	BCBMI	0.08323
CMResAsthma	0.771298	SWB_overall0	0.068102	CMCarArrhythm	2.165607	EQ5DVAS0	0.063433
CMCarHyperten	0.679204	PWB_overall0	0.062277	CMStomInflamm	2.008561	SWB_overall0	0.061488
CMGasPancreas	0.663488	ROL0	0.060521	CMEndHypothy	1.993451	DCEmployment	0.039528
CMEndHypothy	0.542003	DCEmployment	0.045276	CMRheuLupus	1.647547	EQ5DPain0	0.032502
BCBMI	0.506767	EQ5DAnxDep0	0.0362	CMResCOPD	1.407293	DiseaseSite	0.026328
CMRenEndStage	0.498677	DCMarital	0.03589	CMRheuCTD	1.285911	ed_lev	0.025657
EQ5DPain0	0.497696	ed_lev	0.033316	CMCarHeartFail	1.135531	DCMarital	0.025267
EQ5DSelCar0	0.47095	EQ5DPain0	0.032078	CMCarAngina	1.051474	EQ5DAnxDep0	0.023673
CMResCOPD	0.446885	Comorbidities	0.027761	CMNeuParkins	1.035357	EQ5DMob0	0.023085
CMCarVenous	0.422538	PrimaryorMet	0.024138	PWB_overall0	0.911066	Comorbidities	0.021449
CMCarMI	0.40174	StudyArm	0.022059	CMNeuStroke	0.785979	EQ5DSelCar0	0.015177
CMCarArrhythm	0.317881	EQ5DMob0	0.020994	BCBMI	0.7132	StudyArm	0.01505
Comorbidities	0.280834	CMCarHyperten	0.017985	ed_lev	0.622446	PreviousChemo	0.013022
ROL0	0.276591	PreviousChemo	0.013151	EQ5DMob0	0.60776	PrimaryorMet	0.012692
CMEndDiabetes	0.266086	EQ5DSelCar0	0.011003	CMResAsthma	0.573449	CMPrevMal	0.009341
StudyArm	0.258236	CMEndDiabetes	0.010222	EQ5DVAS0	0.556679	CMCarHyperten	0.00828
EQ5DAnxDep0	0.254403	CMPrevMal	0.009266	PreviousChemo	0.515102	CMResAsthma	0.00717
CMStomInflamm	0.24441	CMResAsthma	0.008121	ROL0	0.472313	CMResCOPD	0.005677
EQ5DMob0	0.236114	CMEndHypothy	0.006361	CMEndHyperth	0.433106	CMCarMI	0.005379
CMRheuArth	0.224756	CMSubstAlcohol	0.003922	CMPrevMal	0.405182	CMCarVenous	0.005169
CMCarAngina	0.176192	CMCarMI	0.003742	EQ5DAnxDep0	0.362858	CMEndHypothy	0.004344
EQ5DVAS0	0.164569	CMRheuArth	0.003689	StudyArm	0.336571	CMEndDiabetes	0.004095
CMRheuCTD	0.150602	CMCarVenous	0.003277	SWB_overall0	0.327688	CMCarArrhythm	0.002829
DCEmployment	0.142009	CMResCOPD	0.002494	Comorbidities	0.162286	CMCarAngina	0.00265
CMSubstAlcohol	0.136974	CMStomInflamm	0.002476	DCMarital	0.153898	CMRheuArth	0.002563
DCMarital	0.115902	CMRheuLupus	0.002339	PrimaryorMet	0.135378	CMSubstAlcohol	0.002274
CMNeuParkins	0.104832	CMGasPancreas	0.002232	CMRenEndStage	0.11648	CMStomInflamm	0.001276
CMNeuStroke	0.097995	CMCarAngina	0.001642	CMCarMI	0.109787	CMRheuLupus	0.000868
PrimaryorMet	0.096108	CMRenEndStage	0.001634	CMEndDiabetes	0.098698	CMEndHyperth	0.000723
PreviousChemo	0.082822	CMResChronBron	0.000892	CMStomUlcers	0.091996	CMRheuCTD	0.00058
SWB_overall0	0.071112	CMNeuStroke	0.000826	CMGasPancreas	0.084078	CMRenEndStage	0.000444
CMResChronBron	0.063813	CMCarArrhythm	0.000588	EQ5DSelCar0	0.076894	CMCarHeartFail	0.000305
ed_lev	0.062876	CMEndHyperth	0.00042	DCEmployment	0.063975	CMNeuParkins	0.000191
CMEndHyperth	0.058035	CMRheuCTD	0.000215	CMCarHyperten	0.04455	CMResChronBron	0.000177
PWB_overall0	0.025423	CMNeuParkins	0.000096	EQ5DPain0	0.037622	CMNeuStroke	0.000131
CMGasChronHep	0	CMCarHeartFail	0	DiseaseSite	0.027685	CMStomUlcers	0.000052
CMCarHeartFail	0	CMGasChronHep	0	CMResChronBron	0.000317	CMGasCirrhosis	0
CMResEmphys	0	CMResEmphys	0	CMSubstDrugs	0	CMResEmphys	0
CMRheuPolymyo	0	CMSubstDrugs	0	CMGasCirrhosis	0	CMSubstDrugs	0
CMRheuRhPolymy	0	CMRheuRhPolymy	0	CMGasChronHep	0	CMGasPancreas	0
CMStomUlcers	0	CMGasCirrhosis	0	CMRheuRhPolymy	0	CMRheuPolymyo	0
CMStomMalabsor	0	CMNeuMyasth	0	CMResEmphys	0	CMNeuMyasth	0
CMSubstDrugs	0	CMNeuMS	0	CMRheuPolymyo	0	CMNeuMS	0
CMNeuMS	0	CMStomMalabsor	0	CMStomMalabsor	0	CMGasChronHep	0
CMNeuMyasth	0	CMStomUlcers	0	CMNeuMyasth	0	CMStomMalabsor	0
CMGasCirrhosis	0	CMRheuPolymyo	0	CMNeuMS	0	CMRheuRhPolymy	0
CMRheuPolymyo	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor	0

Table A8: Feature importance ranks for mobility changes at 18 weeks prediction models LR and RF (Chapter 5).

Mobility							
Deterioration				Improvement			
Logistic regression		Random forest		Logistic regression		Random forest	
CMResChronBron	5.681567	FWB_overall0	0.1055	CMEndHyperth	5.327951	EQ5DMob0	0.318964
CMGasPancreas	4.960102	SWB_overall0	0.091013	EQ5DMob0	4.080502	BCBMI	0.078401
CMCarHeartFail	4.945931	PWB_overall0	0.077668	CMCarAngina	3.690014	FWB_overall0	0.070807
CMCarArrhythm	4.130466	EQ5DVAS0	0.073442	CMResCOPD	3.684776	DiseaseSite	0.068345
CMCarMI	3.865736	DiseaseSite	0.051927	CMCarVenous	2.94743	SWB_overall0	0.055995
CMSubstAlcohol	3.775084	DCEmployment	0.048892	CMResChronBron	2.584025	PWB_overall0	0.051819
CMEndHyperth	3.298455	ed_lev	0.040528	CMNeuParkins	2.409078	EQ5DPain0	0.048225
CMResCOPD	3.033808	EQ5DAnxDep0	0.040475	CMCarHyperten	2.35333	EQ5DVAS0	0.042132
CMRheuArth	2.903339	Comorbidities	0.034333	CMResAsthma	2.132524	DCMarital	0.038277
CMRheuLupus	2.42351	DCMarital	0.034039	PreviousChemo	2.022009	EQ5DUsuAct0	0.037502
CMCarVenous	1.366612	EQ5DUsuAct0	0.033967	CMCarMI	1.74107	DCEmployment	0.026439
CMEndDiabetes	1.287484	EQ5DPain0	0.033657	CMEndDiabetes	1.617783	Comorbidities	0.021263
CMResAsthma	1.202195	StudyArm	0.020932	CMEndHypothy	1.571741	EQ5DAnxDep0	0.015872
PreviousChemo	1.0179	PrimaryorMet	0.019977	Comorbidities	1.328023	ed_lev	0.014929
CMNeuParkins	0.761727	EQ5DMob0	0.018281	DiseaseSite	1.312387	CMCarHyperten	0.01356
BCBMI	0.651666	CMResAsthma	0.013156	SWB_overall0	1.307746	PrimaryorMet	0.012237
DiseaseSite	0.57739	PreviousChemo	0.013074	BCBMI	1.165174	PreviousChemo	0.011418
CMPrevMal	0.561657	CMCarHyperten	0.011887	CMCarArrhythm	1.071283	StudyArm	0.00947
PrimaryorMet	0.525226	CMPrevMal	0.010103	CMPrevMal	0.984691	EQ5DSelCar0	0.008962
CMStomInflamm	0.514437	EQ5DSelCar0	0.009966	CMNeuStroke	0.911514	CMPrevMal	0.008718
Comorbidities	0.489691	CMCarArrhythm	0.009579	EQ5DUsuAct0	0.856713	CMEndDiabetes	0.007119
EQ5DPain0	0.463356	CMEndDiabetes	0.009069	FWB_overall0	0.649634	CMCarMI	0.005435
EQ5DMob0	0.433626	CMCarMI	0.008925	CMRheuArth	0.530132	CMResAsthma	0.005372
CMNeuStroke	0.422049	CMRheuArth	0.008617	DCMarital	0.480517	CMEndHypothy	0.005295
EQ5DAnxDep0	0.41443	CMEndHypothy	0.008269	PrimaryorMet	0.457481	CMCarAngina	0.003963
EQ5DSelCar0	0.396671	CMResCOPD	0.003547	StudyArm	0.292456	CMEndHyperth	0.003086
EQ5DVAS0	0.283777	CMCarVenous	0.003416	CMStomInflamm	0.249998	CMCarArrhythm	0.002581
ed_lev	0.263051	CMRheuLupus	0.00247	EQ5DSelCar0	0.224756	CMCarVenous	0.002577
CMCarHyperten	0.117272	CMStomInflamm	0.002402	CMRheuLupus	0.161346	CMResCOPD	0.002568
CMEndHypothy	0.10792	CMNeuStroke	0.002397	CMSubstAlcohol	0.118481	CMRheuArth	0.002414
CMCarAngina	0.087238	CMGasPancreas	0.001793	CMGasPancreas	0.10859	CMResChronBron	0.002178
FWB_overall0	0.0783	CMCarHeartFail	0.001535	CMRenEndStage	0.1067	CMStomInflamm	0.001095
EQ5DUsuAct0	0.073808	CMEndHyperth	0.001213	EQ5DVAS0	0.095971	CMNeuParkins	0.001007
DCMarital	0.057833	CMSubstAlcohol	0.001119	EQ5DAnxDep0	0.092543	CMSubstAlcohol	0.000952
PWB_overall0	0.057653	CMCarAngina	0.000915	ed_lev	0.087568	CMNeuStroke	0.000509
StudyArm	0.048044	CMResChronBron	0.000548	EQ5DPain0	0.075561	CMRheuLupus	0.000224
SWB_overall0	0.039861	CMNeuParkins	0.000193	DCEmployment	0.074864	CMGasPancreas	0.000151
DCEmployment	0.030947	CMResEmphys	0	PWB_overall0	0.00146	CMCarHeartFail	0.00014
CMResEmphys	0	CMSubstDrugs	0	CMRheuCTD	0.000971	CMStomMalabsor	0
CMGasChronHep	0	CMRheuRhPolymy	0	CMCarHeartFail	0.000414	CMSubstDrugs	0
CMGasCirrhosis	0	CMGasChronHep	0	CMStomUlcers	0	CMGasCirrhosis	0
CMStomUlcers	0	CMRheuCTD	0	CMNeuMyasth	0	CMGasChronHep	0
CMRheuPolymyo	0	CMNeuMyasth	0	CMStomMalabsor	0	CMRheuRhPolymy	0
CMRheuRhPolymy	0	CMNeuMS	0	CMGasChronHep	0	CMRheuPolymyo	0
CMRenEndStage	0	CMRenEndStage	0	CMGasCirrhosis	0	CMRheuCTD	0
CMNeuMS	0	CMStomMalabsor	0	CMResEmphys	0	CMResEmphys	0
CMSubstDrugs	0	CMStomUlcers	0	CMRheuRhPolymy	0	CMNeuMyasth	0
CMNeuMyasth	0	CMGasCirrhosis	0	CMRheuPolymyo	0	CMNeuMS	0
CMRheuCTD	0	CMRheuPolymyo	0	CMNeuMS	0	CMRenEndStage	0
CMStomMalabsor	0			CMSubstDrugs	0	CMStomUlcers	0
CMRheuPolymyo	0	CMResEmphys	0	CMNeuMS	0	CMStomMalabsor	0

Table A9: The number of missing data (out of 3260) per column (Chapter 6).

Column name	Number of missing data
SubjectNumber	0
Studystartdate	0
DiseaseSite	0
@_StartTime	0
@_EndTime	0
Dayssincestudystart	0
Weekofcompletion	0
PrimaryFirst	0
Pain	0
Nausea	0
Vomiting	0
Diarrhoea	0
Temperature	0
Chills	0
PhysicalAbility	0
Appetite	0
Fatigue	0
SoreMouth	0
OtherSymptomsYN	0
Indigestion	0
SOB	0
Sex	0
PreviousChemo	0
AgeStudyEntry	0
PrimaryorMet	0
BCBMI	0
Comorbidities	0
CMEndDiabetes	0
DCMarital	20
DCEmployment	36
ed_lev	68
EQ5DUsuAct0	24
EQ5DUsuAct6	157
EQ5DUsuAct12	221
EQ5DUsuAct18	307
stomaYN	0
within_14	0
PWB_overall	9
EQ5DSelCar	10
EQ5DMob	17
EQ5DPain	19
EQ5DAnxDep	26
EQ5DVAS	10
ROL	4
SOC	4
SWB_overall	18
EWB_overall	0
FWB_overall	0
BCCycleDate	0
BCChemoRegimen	0
d_since_cycle	0
Respiratory	0
d_to_event	0
admission_after	0
d_since_previous	0

Table A10: Variable importance os LR and RF predicting admission and admission within 14 days from a symptom report (Chapter 6).

Admission				Admission within 14 days			
Logistic regression		Random forest		Logistic regression		Random Forest	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
stomaYN	1.002035	BCBMI	0.070323	Vomiting	1.120621	Nausea	0.093449
Vomiting	0.825898	Weekofcompletion	0.06977	stomaYN	0.906482	Pain	0.049021
Respiratory	0.732498	AgeStudyEntry	0.058441	PreviousChemo	0.681316	Appetite	0.047587
Temperature	0.648203	EWB_overall	0.047434	Temperature	0.676262	Fatigue	0.047314
PhysicalAbility	0.634755	Nausea	0.041664	OtherSymptomsYN	0.543624	Indigestion	0.043506
Chills	0.60804	DCEmployment	0.038913	PrimaryorMet	0.471119	EQ5DAnxDep	0.043004
SOB	0.482846	FWB_overall	0.038572	Respiratory	0.457451	EWB_overall	0.040581
CMEndDiabetes	0.450962	EQ5DVAS	0.03741	Nausea	0.442808	PhysicalAbility	0.038735
ed_lev	0.39634	SWB_overall	0.036633	SOB	0.424715	OtherSymptomsYN	0.037671
EQ5DPain	0.314148	PWB_overall	0.031502	Diarrhoea	0.392049	Vomiting	0.037661
Diarrhoea	0.303407	ROL	0.031213	Chills	0.387527	Temperature	0.037486
Appetite	0.301007	PhysicalAbility	0.030759	DiseaseSite	0.364315	EQ5DVAS	0.030162
PrimaryorMet	0.261197	d_since_previous	0.027527	PhysicalAbility	0.331507	AgeStudyEntry	0.028406
Pain	0.258675	DiseaseSite	0.027488	Indigestion	0.316126	BCBMI	0.028076
EQ5DUsuAct	0.249282	Fatigue	0.026446	EQ5DUsuAct	0.311672	d_since_previous	0.024331
PreviousChemo	0.241555	Comorbidities	0.025564	EQ5DAnxDep	0.265126	Weekofcompletion	0.02407
DCEmployment	0.240453	OtherSymptomsYN	0.024963	EQ5DMob	0.232594	FWB_overall	0.022663
DiseaseSite	0.235072	Temperature	0.023287	Fatigue	0.213498	Diarrhoea	0.021937
DCMarital	0.226581	Appetite	0.021261	Appetite	0.20659	PWB_overall	0.0219
Indigestion	0.194762	ed_lev	0.021149	DCMarital	0.194407	SWB_overall	0.021796
Weekofcompletion	0.128375	EQ5DAnxDep	0.020025	EQ5DPain	0.136686	SoreMouth	0.02157
SoreMouth	0.119241	Pain	0.019692	Comorbidities	0.110666	DiseaseSite	0.020517
EQ5DSelCar	0.111471	d_since_cycle	0.019465	CMEndDiabetes	0.109877	d_since_cycle	0.018989
Nausea	0.088779	Diarrhoea	0.018937	PWB_overall	0.095847	SOB	0.018566
PWB_overall	0.076334	SOC	0.018463	SoreMouth	0.083617	SOC	0.01795
EQ5DAnxDep	0.060325	SoreMouth	0.018319	Sex	0.078385	ROL	0.016667
OtherSymptomsYN	0.056265	DCMarital	0.017545	EWB_overall	0.05476	EQ5DPain	0.015879
AgeStudyEntry	0.052137	EQ5DPain	0.015706	ed_lev	0.047186	EQ5DUsuAct	0.014584
Sex	0.051936	EQ5DUsuAct	0.015341	BCBMI	0.041283	Comorbidities	0.014396
EQ5DMob	0.047368	Respiratory	0.012351	Weekofcompletion	0.037562	EQ5DMob	0.013554
d_since_previous	0.031228	Indigestion	0.011962	EQ5DVAS	0.035815	ed_lev	0.01256
BCBMI	0.025507	EQ5DMob	0.010798	DCEmployment	0.033648	DCEmployment	0.012327
FWB_overall	0.020703	PrimaryorMet	0.010727	EQ5DSelCar	0.026374	DCMarital	0.011231
d_since_cycle	0.017638	SOB	0.010455	Pain	0.02608	PreviousChemo	0.009691
EQ5DVAS	0.016579	Vomiting	0.009883	d_since_cycle	0.024361	Sex	0.009546
ROL	0.016393	Sex	0.009519	d_since_previous	0.017812	PrimaryorMet	0.008072
Comorbidities	0.009344	PreviousChemo	0.008527	SWB_overall	0.010175	stomaYN	0.007089
EWB_overall	0.005719	Chills	0.00603	FWB_overall	0.007169	Chills	0.005228
SOC	0.002311	CMEndDiabetes	0.005851	SOC	0.0033	CMEndDiabetes	0.004656
SWB_overall	0.001144	EQ5DSelCar	0.005393	ROL	0.001813	EQ5DSelCar	0.004277
Fatigue	0.000807	stomaYN	0.004692	AgeStudyEntry	0.00167	Respiratory	0.003297

Table A11: Variable importance os LR and RF predicting deteriorations in physical well-being and usual activities during chemotherapy (Chapter 6).

Physical well-being				Usual activities			
Logistic regression		Random forest		Logistic regression		Random Forest	
Variable	Importance	Variable	Importance	Variable	Importance	Variable	Importance
stomaYN	1.155651	PWB_overall	0.108305	stomaYN	14.874986	EQ5DUsuAct	0.120806
Respiratory	0.873355	AgeStudyEntry	0.105246	Sex	2.712961	BCBMI	0.11855
Comorbidities	0.643266	BCBMI	0.100533	DiseaseSite	1.744443	PhysicalAbility	0.091943
CMEndDiabetes	0.534641	DiseaseSite	0.057938	EQ5DSelCar	1.297453	AgeStudyEntry	0.054672
PrimaryorMet	0.435864	DCEmployment	0.047136	EQ5DUsuAct	1.23958	Fatigue	0.051943
Vomiting	0.410397	DCMarital	0.041805	Chills	0.899784	DiseaseSite	0.049869
SOB	0.372956	Comorbidities	0.041148	Nausea	0.815692	OtherSymptomsYN	0.032723
Indigestion	0.355405	ed_lev	0.03923	PhysicalAbility	0.763383	EWB_overall	0.029932
DCMarital	0.329574	SWB_overall	0.037837	Respiratory	0.754917	DCMarital	0.029743
PhysicalAbility	0.273179	FWB_overall	0.037627	PrimaryorMet	0.754584	SWB_overall	0.028808
Sex	0.267701	EWB_overall	0.036754	Temperature	0.656508	SoreMouth	0.028684
Appetite	0.264234	EQ5DVAS	0.035539	EQ5DPain	0.613796	SOB	0.025049
EQ5DPain	0.239234	ROL	0.023551	SOB	0.597551	FWB_overall	0.024858
Nausea	0.237095	PreviousChemo	0.022062	Appetite	0.592302	EQ5DVAS	0.024756
ed_lev	0.224421	SOC	0.020822	Vomiting	0.557269	DCEmployment	0.023226
Chills	0.215356	PrimaryorMet	0.020212	Fatigue	0.556349	ed_lev	0.021449
EQ5DMob	0.207578	Sex	0.019125	CMEndDiabetes	0.540794	PWB_overall	0.018827
Pain	0.174175	EQ5DAnxDep	0.015189	PreviousChemo	0.52831	EQ5DAnxDep	0.018638
OtherSymptomsYN	0.148012	d_since_cycle	0.014394	OtherSymptomsYN	0.520016	EQ5DMob	0.018047
PreviousChemo	0.110108	EQ5DPain	0.013797	EQ5DAnxDep	0.471259	Sex	0.015929
EQ5DUsuAct	0.098296	EQ5DMob	0.012945	SoreMouth	0.427144	ROL	0.014277
Fatigue	0.090866	EQ5DUsuAct	0.012776	Comorbidities	0.409742	Comorbidities	0.013579
PWB_overall	0.067339	SoreMouth	0.012368	Pain	0.368626	SOC	0.012519
EQ5DAnxDep	0.052104	d_since_previous	0.012338	ed_lev	0.349721	EQ5DSelCar	0.012447
EWB_overall	0.040547	OtherSymptomsYN	0.010555	DCMarital	0.240056	Appetite	0.012098
Temperature	0.037263	stomaYN	0.010229	Diarrhoea	0.153736	Chills	0.011685
DCEmployment	0.028061	Respiratory	0.009964	EQ5DMob	0.121249	Nausea	0.011138
d_since_cycle	0.023004	Appetite	0.009377	SWB_overall	0.103758	Pain	0.010743
Diarrhoea	0.019887	PhysicalAbility	0.009113	EWB_overall	0.095425	EQ5DPain	0.01058
DiseaseSite	0.016585	Nausea	0.00831	d_since_previous	0.092215	d_since_previous	0.010049
EQ5DSelCar	0.016274	Pain	0.007878	BCBMI	0.055711	Respiratory	0.007296
FWB_overall	0.015914	CMEndDiabetes	0.007778	DCEmployment	0.048918	PrimaryorMet	0.007276
SoreMouth	0.012704	Fatigue	0.007362	AgeStudyEntry	0.037853	Diarrhoea	0.007226
d_since_previous	0.012615	Diarrhoea	0.006731	PWB_overall	0.02852	d_since_cycle	0.007027
SOC	0.012511	SOB	0.00602	EQ5DVAS	0.025398	stomaYN	0.006444
BCBMI	0.009855	Indigestion	0.004972	d_since_cycle	0.023839	Indigestion	0.004586
SWB_overall	0.009779	Vomiting	0.004776	SOC	0.011694	PreviousChemo	0.004527
ROL	0.004439	Temperature	0.003533	FWB_overall	0.011616	Temperature	0.003033
AgeStudyEntry	0.003416	EQ5DSelCar	0.003052	ROL	0.0011	CMEndDiabetes	0.002783
EQ5DVAS	0.003186	Chills	0.001671	Indigestion	0.000489	Vomiting	0.002235