Computer Methods and Programs in Biomedicine 263 (2025) 108653

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Towards clinical prediction with transparency: An explainable AI approach to survival modelling in residential aged care

Teo Susnjak *, Elise Griffin

School of Mathematical and Computational Sciences, Massey University, Albany, Auckland, 0632, New Zealand

ARTICLE INFO

Keywords:
Explainable AI (XAI) in geriatric palliative care
Survival analysis machine learning aged care mortality prediction
Explainable clinical decision support systems for end-of-life care
Time-to-event analysis

ABSTRACT

Background and Objective: Scalable, flexible and highly interpretable tools for predicting mortality in residential aged care facilities for the purpose of informing and optimizing palliative care decisions, do not exist. This study is the first and most comprehensive work applying machine learning to address this need while seeking to offer a transformative approach to integrating AI into palliative care decision-making. The objective is to predict survival in elderly individuals six months post-admission to residential aged care facilities with patient-level interpretability for transparency and support for clinical decision-making for palliative care options.

Methods: Data from 11,944 residents across 40 facilities, with a novel combination of 18 features was used to develop predictive models, comparing standard approaches like Cox Proportional Hazards, Ridge and Lasso Regression with machine learning algorithms, Gradient Boosting (GB) and Random Survival Forest. Model calibration was performed together with ROC and a suite of evaluation metrics to analyze results. Explainable AI (XAI) tools were used to demonstrate both the cohort-level and patient-level model interpretability to enable transparency in the clinical usage of the models. TRIPOD reporting guidelines were followed, with model parameters and code provided publicly.

Results: GB was the top performer with a Dynamic AUROC of 0.746 and a Concordance Index of 0.716 for six-month survival prediction. Explainable AI tools provided insights into key features such as comorbidities, cognitive impairment, and nutritional status, revealing their impact on survival outcomes and interactions that inform clinical decision-making. The calibrated model showed near-optimal performance with adjustable clinically relevant thresholds. The integration of XAI tools proved effective in enhancing the transparency and trustworthiness of predictions, offering actionable insights that support informed and ethically responsible end-of-life (EoL) care decisions in aged care settings.

Conclusion: This study successfully applied machine learning to create viable survival models for aged care residents, demonstrating their usability for clinical settings via a suite of interpretable tools. The findings support the introduction into clinical trials of machine learning with explainable AI tools in geriatric medicine for mortality prediction to enhance the quality of EoL care and informed discussions regarding palliative care.

1. Introduction

Predicting death is easy. Everybody will die. Estimating the precise probability of death for an individual within a specific time period is more difficult. An accurate estimate of expected survival time helps people choose treatments that align with their goals of care [1]. A falsely optimistic prognosis reduces the quality of death experienced by a patient and their loved ones [2].

Palliative care in people with a terminal diagnosis focuses on withdrawing treatments that cause pain or suffering and offering care that enhances the quality of remaining life. Many people are willing to endure short-term discomfort to increase survival time but there comes a point when sacrificing quality for quantity is no longer justifiable. For some, this realization comes just days before death, while for others, it may be recognized several months prior [3].

People entering residential aged care do not usually have a specific terminal illness. Rather, they are undergoing the inexorable decline in function that accompanies chronic illness and natural ageing [4]. Shifting from an active treatment model to a palliative approach in this setting is a nuanced decision and is not always clearly communicated with residents or their families [5]. However, more than one-third

* Corresponding author.
E-mail address: t.susnjak@massey.ac.nz (T. Susnjak).

https://doi.org/10.1016/j.cmpb.2025.108653
Received 1 September 2024; Received in revised form 21 January 2025; Accepted 5 February 2025
Available online 15 February 2025

of older people admitted to residential aged care will die within six months of admission [6]. In most healthcare settings, this prognosis would prompt discussions about the patient's care preferences in view of their short life expectancy. Yet these vital conversations occur less frequently than many older people would prefer [7].

In gerontological research, prognostic tools in mortality prediction with the potential to inform palliative care options can be broadly categorized into prognostic indices, statistical methods, and machine learning approaches. Prognostic indices aggregate multiple clinical variables correlated with mortality into a single value, and have been the most popular tool for predicting mortality in aged care contexts [8]. These indices, particularly those utilizing the Minimum Data Set (MDS), have played a pivotal role in predicting mortality in residential aged care. Notable tools such as the MDS Mortality Risk Index (MMRI) and its various adaptations [9–13] have provided valuable insights into mortality risks. However, new technologies have emerged which can address many of the shortcomings of earlier approaches. For instance, machine learning methods offer the potential for greater accuracy and the capacity to manage complex, non-linear data relationships [14]. The application of machine learning in medical contexts has increased rapidly, with some hospitals already integrating this technology into their everyday decision-making processes [15].

Palliative care has yet to significantly benefit from advancements in machine learning, despite the recognized potential of these technologies to enhance the quality of end-of-life (EoL) care by facilitating timely and informed palliative treatment decisions [16]. Current applications of machine learning for mortality prediction and palliative care decision-making are predominantly confined to hospital environments and terminally ill patients [16], leaving a notable gap in their application to residential aged care settings. This gap is particularly significant given the need for predictive tools that address the complexities of aged care populations while ensuring transparency and clinical usability. Furthermore, the lack of explainability behind many machine learning models' predictions, presents a major barrier to their adoption in healthcare. The opaque nature of these models often obscures the reasoning behind their predictions, undermining clinician trust and impeding their integration into decision-making processes [15]. To address these challenges, recent advancements in eXplainable AI (XAI) techniques, such as SHAP and LIME, have introduced methods to enhance the transparency of machine learning models [17–19]. However, their application in mortality prediction for residential aged care remains unexplored. This study fills this critical gap by combining machine learning with XAI tools to predict six-month mortality in aged care residents, delivering a transparent and actionable framework for supporting palliative care decision-making in this underserved context.

Study contributions and novelty

The contributions of this study are as follows:

1. This work is the first comprehensive study to develop machine learning models for predicting survival probabilities for a general population of adults in residential aged care facilities.
2. This research is novel in that it integrates machine learning into supporting palliative care, tailored specifically for aged care settings, addressing existing gaps in the literature [15,16] by providing actionable insights that can improve the timing and quality of EoL care decisions.
3. This study pioneers the use of XAI tools in survival analysis models specifically designed for residential aged care, offering both *cohort-level* model interpretability and *individual-level* prediction explanations to enhance overall transparency and trust in AI-driven clinical decision-making where current literature shows a significant lack of XAI integration in survival analysis models [20].
4. Our extensive comparative benchmarking analysis demonstrates the superior performance and interpretability of machine learning models over state-of-the-art statistical models and prognostic indices, thus validating its practical relevance in real-world aged care settings.
5. Our study leverages both a unique combination of features and a large dataset to develop models capable of supporting personalized care, which aligns with the latest trends in biomedical AI research.
6. Our workflow and the set of technologies used significantly advance EoL care in aged care settings, enabling more informed, transparent, and ethically responsible decision-making that aligns with patient preferences, thus bridging the gap between theoretical models and real-world clinical utility [15].

2. Methods

This study is reported according to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guideline [21].

2.1. Data source and participants

Healthcare data used in this study were collected during the provision of routine care to older individuals admitted for long-term care between 1st July 2017 and 30th August 2023 to facilities owned by a single large Australasian private residential aged care provider. Data from residents at 34 New Zealand and 6 Australian aged care facilities were included. Individuals were eligible for inclusion if they were admitted for long-term care on or after 1st July 2017.

2.2. Outcome variable and problem formulation

The primary model prediction outcomes were twofold:

1. The development of a survival function $S(t|x)$, where t represents the time since admission to an aged care facility and x denotes the set of covariates or features. This function provides a continuous probability curve of survival from admission up to six years, offering a comprehensive longitudinal risk assessment, that also enables relative risk comparisons between patients.
2. The generation of calibrated, time-specific survival probabilities at the key intervals $t = 1, 3, 6$, and 12 months post-admission. These intervals represent time points for clinical decision-making, with the 6-month forecast being particularly vital in the context of this study.

For each individual i in our dataset D, consisting of n subjects, we observe the time-to-event T_i and an event indicator δ_i, where $\delta_i = 1$ denotes the event occurrence (death) and $\delta_i = 0$ indicates censoring (survival). The dataset is represented as $D = \{(T_1, \delta_1, x_1), \ldots, (T_n, \delta_n, x_n)\}$, where x_i are the features. The trained model generates a survival probability $S(t|x_i)$ for each patient i at any given time t, effectively quantifying the risk of the event (death) across the entire timeframe.

Post-training, model calibration is a necessary step in survival analysis that aligns the above-predicted probabilities with actual observed outcomes, ensuring the reliability and accuracy of the predictions for a specific point in time of interest. To calibrate the survival probabilities predicted by our model at any time point t, we used Platt scaling [22,23]. This involved fitting a logistic regression model to the survival model's output, where $P(T > t|x)$ represents the calibrated survival probability beyond time t, given features x. The calibration is formally expressed as:

$$P(T > t|x) = \frac{1}{1 + \exp(A \cdot f(x,t) + B)}, \quad (1)$$

where A and B are parameters derived from logistic regression, and $f(x,t)$ is the survival model's uncalibrated prediction at time t. The model thus refines the initial predictions $f(x,t)$ using the features x and the specific time point t, enhancing the alignment of predicted survival probabilities with observed survival rates. This calibration step is regarded as integral to transforming statistical predictions into clinically actionable insights, particularly for time-sensitive decision-making.

2.3. Features

A total of 18 features were used for modelling, drawn from demographic and clinical data recorded by registered nurses during the initial clinical evaluation of newly admitted residents. Medication data were taken from the electronic medicine chart. The earliest instance of each specific feature following admission was used. Any feature recorded more than 31 days after admission was excluded since our objective was to predict survival probability from the time of admission. Therefore, data collected more than one month post-admission was discarded.

Tables 1 and 2 list all features included in the model and the values assigned to each level. Table 1 details the demographic attributes of the study cohort, including reasons for discharge and Rx-Risk Co-morbidity Index diagnostic categories [24]. Table 2 reports clinical variables drawn from the initial nursing assessment. Domain expertise was used to assign ordinal values representing the degree of severity for each level of the features. High values represent a state associated with worse function or clinical status (higher mortality risk) and low values represent states associated with better function or clinical status (lower mortality risk).

Features relating to nutrition, mobility, smoking, sleep, skin integrity and continence correspond to the standardized set of questions and answers based on the InterRAI long-term care facilities assessment [25]. Features relating to current health status, cognition, mood and pressure ulcers are drawn from validated InterRAI-based composite measures (respectively, the CHESS scale (Changes in Health, End-Stage Disease, Symptoms and Signs) [26,27], cognitive performance scale [28], depression rating scale [29], pressure ulcer risk scale [30]). Co-morbidities are assessed in two ways. The presence of a specific diagnosis was established by filtering the diagnosis fields in the resident clinical record for terms that captured dementia of any type, ischemic cardiac disease and heart failure of any type, any malignant neoplasm, any non-cancer pulmonary disease and any form of diabetes, excluding glucose intolerance. The sum of the scores for these items (1 = ANY diagnosis of this type present, 0 = ALL diagnoses of this type absent), rather than the individual diagnosis, was used as a feature in the final model (minimum value = 0, maximum value =5). The Rx-Risk Co-morbidity Index was included as a second comorbidity item. This index provides a weighted score for each specific diagnosis based on prescription data. The scores are summed for an individual resident to provide the final Rx-Risk score. The falls feature was a bespoke question used by the provider about the frequency of falls in the past six months prior to admission.

During the preliminary stages of our data processing, a pairwise correlation coefficient threshold of 0.7 was used as a guide for eliminating highly correlated variables to ensure model parsimony and reduce multicollinearity [31]. The decision on which one of the variables to exclude from the model was made by examining data quality and completeness, the relative univariate predictive power of the variable, and potential interpretability.

2.4. Dataset size and missing data

Sample size was determined by data availability. Complete digital personal health records for all residents and electronic medicine chart data were available from July 1st 2017. We used all data from current and discharged long-term residents admitted on or after this date for model development. Missing data was encountered at varying degrees for most features, as reported in Table 2. Features with 75% or more missing values were excluded. The Multiple Imputation by Chained Equations (MICE) [32] was used to impute missing values for all features, in line with similar survival analysis studies [33,34].

Table 1
Cohort demographic features, their frequencies and transformation values used for modelling.

Category	Residents (n)	Residents (%)	Value
Age (years)			
65–69	248	2%	67
70–74	645	5%	72
75–79	1435	12%	77
80–84	2445	20%	82
85–89	3255	27%	87
90–94	2725	23%	92
95–99	1060	9%	97
100+	130	1%	100
Age (mean,SD)	85.7 (mean)	7.2 (SD)	
Gender			
Female	7200	60%	0
Male	4494	38%	1
Other/Gender Diverse	167	1%	0
Unknown	82	1%	0
Discharge Reason			
Deceased	6725	56%	1
Current resident	3465	29%	0
Transfer to another care facility	1145	10%	0
Discharged home	351	3%	0
Transfer to public hospital	244	2%	0
Transfer to hospice	<50	<1%	1
Rx-Risk Comorbidity Index[24]			
Pain	7151	79%	3
Psychotic disorder	3080	34%	6
Congestive heart failure	2383	26%	2
Gastrooesophageal reflux disease	2213	24%	0
Ischemic heart disease	1751	19%	−1
Depression	1693	19%	2
Antiplatelets	1449	16%	2
Anticoagulants	1132	12%	1
Hyperlipidaemia	1030	11%	−1
Anxiety	958	11%	1
Chronic airways disease	932	10%	2
Allergies	772	9%	−1
Steroid-responsive disease	757	8%	2
Diabetes	636	7%	2
Ischemic heart disease	581	6%	2
Hypertension	540	6%	−1
Dementia	498	5%	2
Glaucoma	469	5%	0
Hypothyroidism	430	5%	0
Gout	403	4%	1
Arrhythmia	325	4%	2
Malignancies	325	4%	2
Osteoporosis/Pagets	321	4%	−1
Inflammation/pain	308	3%	−1
Parkinsons disease	306	3%	3
Benign prostatic hypertrophy	279	3%	0
Incontinence	247	3%	0
Epilepsy	246	3%	0
Benign prostatic hyperplasia	200	2%	0
Renal disease	59	1%	6
Smoking cessation	59	1%	6

2.5. Predictive algorithms

This study explored traditional algorithms[1] comprising Lasso Regression, Ridge Regression (RR) and Cox Proportional Hazards (CPH) alongside the following machine learning approaches:

[1] The implementations of the algorithms from the Python library scikit-survival [35] version 0.21.0 were used and XGBoost [36] version 1.7.6.

Table 2
List of features, together with their raw values, the assessment used to capture them, their distribution as well as their transformation for modelling.

Features	Assessment question	Answer	Residents (n)	Residents (%)	Value
Falls	History of falls	No history of falls	5125	42.9%	0
		4 or less in last 6 months	5270	44.1%	1
		5 or more in last 6 months	477	4.0%	2
		3 or more falls in one month period	306	2.6%	3
		Missing	766	6.4%	
Health status	What was the CHESS scale score?	No symptoms	1446	12.1%	0
		Minimal health instability	1512	12.7%	1
		Low health instability	1444	12.1%	2
		Moderate health instability	793	6.6%	3
		High health instability	385	3.2%	4
		Highest level of instability	65	0.5%	5
		Missing	6299	52.7%	
Comorbidities	What was the weighted Rx-Risk scale score?	Sum of weighted scores (range −3 to 23)	9065	75.9%	
		Missing	2879	24.1%	
	Was this diagnosis present?	Dementia	5179	43.4%	1
		Heart disease	3658	30.6%	1
		Cancer	1301	10.9%	1
		Diabetes	1289	10.8%	1
		Lung disease	1112	9.3%	1
Cognition	What was the cognitive performance scale score?	Intact	639	5.3%	0
		Borderline intact	538	4.5%	1
		Mild impairment	2127	17.8%	2
		Moderate impairment	1527	12.8%	3
		Moderate/Severe impairment	199	1.7%	4
		Severe impairment	426	3.6%	5
		Very severe impairment	99	0.8%	6
		Missing	6389	53.5%	
Mood	What was the depression rating scale score?	None (0)	2609	21.8%	0
		Mild (1–2)	1729	14.5%	1
		Moderate (3–5)	909	7.6%	2
		Severe (6–14)	308	2.6%	3
		Missing	6389	53.5%	
Nutrition	Has the resident lost weight recently?	No	4379	36.7%	0
		Unsure	5242	43.9%	0
		Yes	1650	13.8%	1
		Missing	673	5.6%	
	Is the resident eating poorly or has a lack of appetite?	No	9136	76.5%	0
		Yes	2135	17.9%	1
		Missing	673	5.6%	
Mobility	How does your resident mobilize?	Independent	4175	34.9%	0
		Supervision or prompting	1889	15.8%	1
		1 person assistance	2395	20.1%	2
		2 person assistance	921	7.7%	3
		Does not mobilize (bed or chair bound)	1153	9.7%	4
		Missing	1411	11.8%	
	What equipment does your resident use to mobilize safely?	None	2698	22.6%	0
		Walking stick	813	6.8%	1
		Walking frame	5083	42.6%	2
		Transfer belt or other	586	1.7%	3
		Gutter frame	220	1.8%	4
		Wheelchair, fallout chair or lazyboy	1130	9.5%	5
		Missing	1411	11.8%	
Smoking	Has your resident smoked in the past?	No	7604	63.7%	0
		Yes	1950	16.3%	1
		Missing	2390	20.0%	
Skin	Has your resident's skin integrity changed since last assessment?	Improved	170	1.4%	0
		No Change	2609	21.8%	0
		Fluctuated	157	1.3%	1
		Declined	530	4.4%	2
		Missing	8478	71.0%	
	What was the Pressure Ulcer Risk scale?	Very low risk	2629	22.0%	0
		Low risk	1967	16.5%	1
		Moderate risk	554	4.6%	2
		High risk	349	2.9%	3
		Very high risk	34	0.3%	4
		Missing	6411	53.7%	
Continence	Is the resident incontinent of faeces?	No	3093	25.9%	0
		Yes	2187	18.3%	1
		Missing	6664	55.8%	
	Is the resident incontinent of urine?	No	3808	31.9%	0
		Yes	1554	13.0%	1
		Missing	6582	55.1%	

Random survival forest (RSF). Extends the Random Forest algorithm to survival data by constructing an ensemble of survival trees. Each tree is grown using a bootstrap sample of the data, and the cumulative hazard function for an individual is estimated by averaging the Nelson–Aalen

Table 3
Overview of hyperparameter settings for the models.

Model	Hyperparameters		
CPH	None		
RR	l1_ratio=10^{-100}	n_alphas=1	alphas=2.24e−06
	normalize=True	fit_baseline_model=True	
Lasso	l1_ratio=0.9	alpha_min_ratio=0.01	fit_baseline_model=True
GBM	n_estimators=771	min_samples_split=20.04	max_depth=7
	min_samples_leaf=1.85	learning_rate=0.28	dropout_rate=0.05
	objective='survival:cox'	max_features=4	subsample=0.83
XGB	num_boost_round=1107	learning_rate=0.018	max_depth=3
	colsample_bytree=0.83	gamma=0.49	objective='survival:cox'
	subsample=0.58		
RSF	n_estimators=592	min_samples_split=2.54	max_depth=7
	min_samples_leaf=20.89		

estimators from the terminal nodes of all trees. The RSF does not rely on the proportional hazards assumption, making it flexible for capturing complex interactions between features and time-to-event data.

Gradient boosting machine (GBM). Conceptualized by [37], employs an optimization strategy for a differentiable loss function L in survival analysis. This is achieved through the iterative addition of regression trees, with each tree addressing the residuals of the model up to that point. In each iteration t, GBM adapts the model by adding a new regression tree, specifically targeting the negative gradients (denoted as $-\nabla L$) of the loss function—reminiscent of the steepest descent method in optimization. The ensemble model is updated according to:

$$F_t(x) = F_{t-1}(x) + \nu \cdot \text{Tree}(x, -\nabla L), \tag{2}$$

where $F_t(x)$ represents the model's prediction at iteration t, $F_{t-1}(x)$ is the prediction from the previous iteration, ν is a shrinkage parameter introducing regularization to control model complexity, and $\text{Tree}(x, -\nabla L)$ signifies the regression tree added in the current iteration, tailored to the negative gradient of the loss. This methodical enhancement at each step incrementally refines the model, contributing to a more robust and accurate predictive performance.

XGBoost (XGB). Enhances the GBM algorithm [38] by integrating an additional regularization term Ω into its optimization process, aiming to reduce overfitting by penalizing model complexity. This model adopts a sophisticated tree learning algorithm, leveraging second-order gradient statistics, which facilitates more precise split decisions during tree construction. The objective function of XGB at iteration t is formulated as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{3}$$

where l represents the loss function, $\hat{y}_i^{(t-1)}$ denotes the prediction at the previous iteration, and f_t symbolizes the tree structure added in the current iteration.

2.6. Experimental setup

Standardization. Before running experiments with the above algorithms, data was standardized for the Lasso and RR algorithms.

Hyperparameter tuning. Tuning of hyperparameters was executed using random train/test splits. This process involved exploring a range of hyperparameters, guided by the goal of maximizing model performance[2]. Table 3 summarizes the resulting hyperparameters used in the subsequent experiments for each algorithm, reflecting the settings that yielded the best results in our analyses.

[2] Interested readers can refer to the GitHub repository https://github.com/teosusnjak/survival-analysis-stage1 for complete implementation details

Training and test data setup. Subsequent to hyperparameter tuning, each algorithm was tested in 20 experiments using different train/test splits at a 90/10 ratio. For algorithms requiring a validation set, the training set was further divided using a 90/10 split. Outcomes from these experiments were aggregated and presented with 95% confidence intervals.

Model evaluation metrics. Model performance was assessed using various evaluation measures. The Concordance Index (C-index), as per [39], assesses the model's ability to rank survival times.

$$C - index = \frac{\sum_{i:\delta_i=1} \sum_{j:T_j>T_i} \mathbb{I}(\hat{y}_i > \hat{y}_j)}{\sum_{i:\delta_i=1} \sum_{j:T_j>T_i} 1} \tag{4}$$

where T_i, T_j are observed survival times, $\hat{y}_i$, $\hat{y}_j$ are predicted times, and δ_i indicates uncensored events. Meanwhile, Harrell's C-index [40] extends the above to account for censored data, providing a robust evaluation in complex risk scenarios. Dynamic AUROC [41], evaluates the model's time-specific discriminative power, distinguishing individuals based on event occurrence at predetermined time points, defined as:

$$AUROC(t) = \int Sensitivity(t, c) \times [1 - Specificity(t, c)]\, dc \tag{5}$$

with AUROC computed as a function of time t, integrating sensitivity and specificity over thresholds c. While he Integrated Brier Score (IBS) [42] measures both calibration and discrimination, offering an average prediction error for survival probabilities:

$$IBS = \frac{1}{T} \int_0^T [1 - n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i(t))^2]\, dt \tag{6}$$

integrating mean squared difference between observed outcomes y_i and predictions $\hat{y}_i(t)$ over period T, adjusted for censoring. Additionally, calibrated ROC Analysis, following [43], was utilized to analyse the model's specificity and sensitivity, revealing predictive power variations at different thresholds. The Hosmer–Lemeshow test [44] was used to gauge model calibration by comparing observed and predicted outcomes, with lower chi-squared values indicating better calibration.

Interpretation of evaluation metrics. The C-index measures the model's ability to rank survival times correctly. A value of 1.0 indicates perfect discrimination, meaning the model perfectly ranks all survival times, while 0.5 reflects random performance. Values above 0.7 are typically considered indicative of strong predictive ability. Harrell's C-index can be interpreted on the same basis, but it extends this metric to account for censored data, making it robust for incomplete survival information. Meanwhile, the Dynamic AUROC evaluates the model's ability to distinguish between individuals experiencing events at specific time points. Values close to 1.0 indicate excellent discrimination, while values below 0.7 suggest limited utility. This metric helps track the model's reliability over time-specific predictions. IBS on the other

hand combines discrimination and calibration by quantifying average prediction error. Scores closer to 0 indicate better predictive accuracy, with ideal values typically below 0.25 in well-performing survival models. Calibrated ROC analysis examines the balance between sensitivity (true positive rate) and specificity (true negative rate) across different thresholds. High sensitivity and specificity reflect the model's ability to identify outcomes accurately. Finally, the Hosmer–Lemeshow test evaluates calibration quality, with lower χ^2 values and higher p-values (e.g., > 0.05) indicating good agreement between observed and predicted outcomes.

Model calibration. The multi-metric approach enables the evaluation of both uncalibrated and time-specific metrics and offers a comprehensive understanding of each model's predictive accuracy, discriminative power, and reliability. The above metrics serve both as individual model performance indicators and as tools for model comparison. The best-performing model was subsequently calibrated for time-specific predictions at a six-month time point using Platt scaling, discussed earlier. The effectiveness of the calibration also entailed performing 20 train/test splits at a 90/10 ratio to evaluate the accuracy, which was visualized via a calibration plot and reported through Dynamic AUROC, IBS, the standard Concordance Index (C-index) and Harrell C-index. The specificity, sensitivity and negative predictive power (accuracy in forecasting mortality for the people who died within six months of admission) of this model were inspected via ROC curve analyses. The final best-performing model from the analyses, its calibration together with an example code on how to use it, is publicly available from a GitHub repository.[3]

2.7. eXplainable AI tools

Machine learning models are often categorized based on their transparency into interpretable or "black-box" models. Interpretable models, such as Linear Regression and Decision Trees, are inherently understandable without additional tools, allowing their global behaviour (i.e. across the entire dataset) to be directly interpreted. Black-box models, such as Gradient Boosting and Neural Networks, require *post-hoc* methods for understanding their mechanics. In this study, we define the term *interpretability* as understanding the global behaviour of a model, which includes how input features and their values influence predictions and which features are most important in driving the model's decisions. *Explainability*, we define as understanding instance-specific (i.e. patient-level) predictions using post-hoc methods like SHAP. *Model transparency* captures both interpretability of its mechanics and the explainability of its individual predictions.[4]

Balancing predictive strength and model transparency is a specific challenge in contemporary machine learning research, especially in sensitive fields such as healthcare where decisions based on model outputs must adhere to ethical and regulatory standards. XAI aims to address this challenge by enhancing the transparency of machine learning models. Among various XAI tools, SHapley Additive exPlanations (SHAP) is prominent for its effectiveness in elucidating the decision-making processes of complex models [45]. SHAP provides both the model interpretability at the global (cohort-level) and explainability at the local (patient-level) dimensions. At the cohort-level, SHAP Summary plots depict the ranked impact of features on survival probability, while SHAP Dependence plots reveal interactions between features. These plots offer a macroscopic view of the key factors influencing model predictions. On the patient-level, SHAP Waterfall plots detail each feature's contribution to a specific prognosis, enhancing the model's explainability and thereby, clinician trust in the model's outputs. It is especially in this dimension that these tools exceed the capabilities of traditional prognostic indices and statistical models. Formally, SHAP values are calculated using cooperative game theory principles, quantifying the average marginal contribution of a feature across all possible combinations. The SHAP value for feature j in prediction instance i is given by:

$$SHAP_{ij} = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(x_{i,S \cup \{j\}}) - f_S(x_{i,S}) \right] \quad (7)$$

Here, F represents the set of all features, S is a subset of F excluding feature j, $|S|$ and $|F|$ denote the number of features in S and F respectively, $f_{S \cup \{j\}}(x_{i,S \cup \{j\}})$ is the model prediction with feature j, and $f_S(x_{i,S})$ is the prediction without feature j. The equation effectively decomposes a prediction into the sum of contributions from each feature, providing an interpretable perspective of how each feature impacts the model's prediction.

3. Results

3.1. Overview

The study's results are presented in two sections: The first part details the performance of uncalibrated models predicting survival probability up to six years post-admission, including insights from individualized survival curves and SHAP plots. The second part focuses on a calibrated model's performance at the six-month mark, complemented by ROC analyses.

3.2. Final study participants

Data from 12882 individuals were extracted from the database. 407 individuals lacked requisite assessment data within 31 days of admission from the cohort and were removed. Reconciling duplicate data from residents with one or more consecutive admissions resulted in 11945 unique individuals. One resident was excluded due to a negative value for length of stay. The final study cohort comprised 11944 individuals. The mean age in the cohort was 86 years (SD=7), with the majority being women (n=7200, 60%). Just over half the cohort (n=6739, 56%) were discharged due to death (the modelled outcome) and approximately 30% were current residents (n=3465). When all categories were combined, in total 57% (n=6740) were represented deceased and 43% (n=5205) comprised censored (survived) (see Table 1). Three-quarters of residents had an electronic medicine chart initiated within 31 days of admission, enabling the estimation of the Rx-Risk Comorbidity Index for these individuals (n=9065, 76%) [24]. The most common diagnostic categories based on prescription data were pain, a psychotic disorder (most likely behavioural and psychological symptoms of dementia), congestive heart failure and gastro-oesophageal reflux. The distribution of Rx-Risk Comorbidity Index categories is reported[5] in Table 1.

3.3. Model performance

Table 4 shows evaluation results of survival models across a time horizon of up to 74 months. The best-performing models, according to the C-index, are the ensemble methods, GBM, RSF and XGB with negligible differences between them. These models exhibit C-indices of 0.712 to 0.714, supported by narrow 95% confidence intervals, indicating effective discriminatory power and robust statistical stability. The leading C-index of the top three models is complemented by their Harrell's C-index score of ~0.67 and an AUROC of ~0.75, confirming effective performance in both discrimination and calibration. Only marginally lower, CPH, RR and Lasso regression exhibit similar performance on this dataset across all the metrics.

[3] https://github.com/teosusnjak/survival-analysis-stage1

[4] While interpretability and explainability are often used interchangeably in the literature, this study adopts a clear distinction since local and global aspects of model behaviour are crucial to this study where this distinction is particularly relevant for clinical applications.

[5] For brevity, less common diseases with an incidence of fewer than 50 patients are omitted from the table.

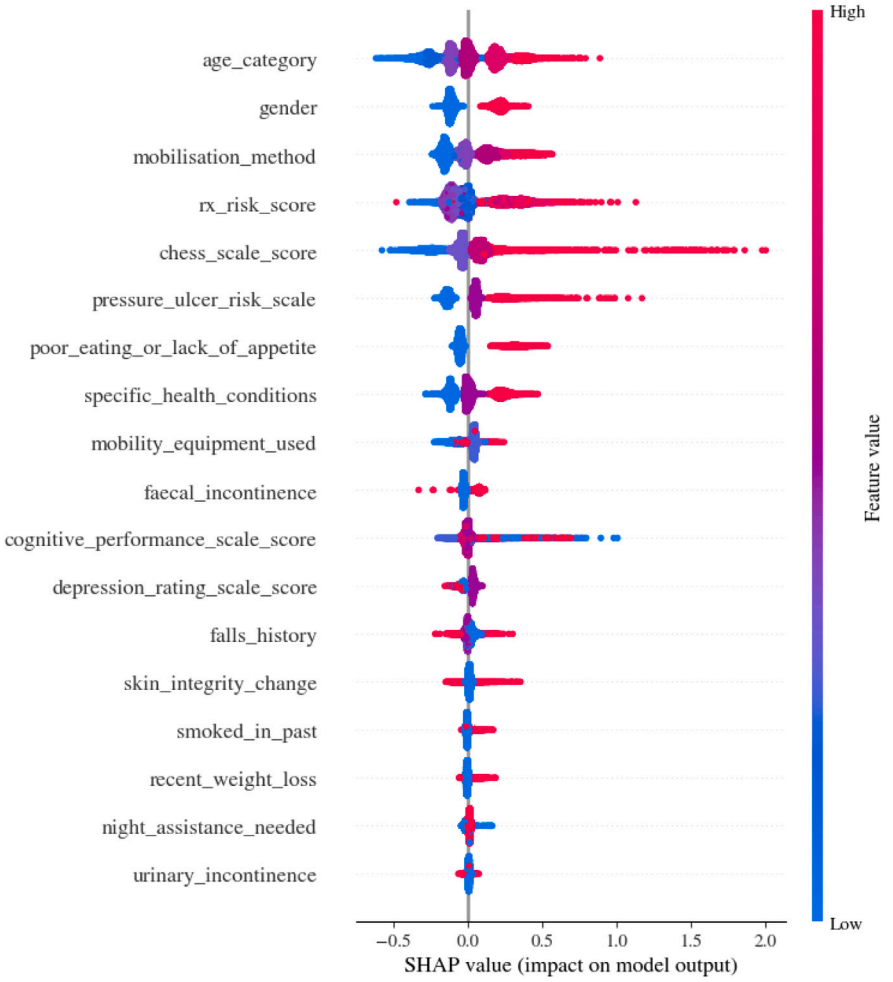


Fig. 1. Feature importance summary plot for the XGB model.

Table 4
Rank-ordered performance metrics of different models across the entire survival period up to 74 months post-admission to a care facility with 95% confidence intervals.

Model	C-index	Harrell's C-index	AUROC
GBM	0.714 (0.711–0.717)	0.673 (0.67–0.676)	0.747 (0.743–0.751)
RSF	0.712 (0.708–0.715)	0.671 (0.667–0.674)	0.745 (0.739–0.75)
XGB	0.712 (0.709–0.716)	0.675 (0.672–0.679)	0.755 (0.75–0.759)
CPH	0.709 (0.706–0.711)	0.671 (0.668–0.674)	0.749 (0.746–0.752)
RR	0.708 (0.705–0.711)	0.670 (0.667–0.673)	0.745 (0.742–0.749)
Lasso	0.706 (0.704–0.709)	0.666 (0.663–0.669)	0.742 (0.739–0.746)

Model interpretability and explainability

Here we examine the internal mechanics of the model and the impact of each feature at the cohort-level and at the individual patient-level. For this analysis, we selected XGB,[6] one of the top performing models according to Table 4, for a more detailed inspection. The SHAP summary plot in Fig. 1 presents the features included in the model ranked in order of importance (features with the greatest influence on predicted survival at the top). The SHAP values are shown on the x-axis. SHAP values quantify the contribution of each feature to the model estimate of mortality risk, in deviation from the mean prediction. The grey vertical line represents a zero-impact mean prediction. Positive values (to the right of the zero-impact line) are associated with increased mortality risk and negative values (to the left of the zero-impact line) with reduced mortality risk. As the data points for each feature move further from the vertical line, the greater the impact of this feature on expected survival becomes. The colour spectrum (blue to red) across the SHAP value scatter shows the value of the feature value, with blue indicating lower values and red signifying higher values. This relationship is most easily visualized in the 'age_category' and 'chess_scale_score', where higher values (older age or worse health status respectively) are both red and associated with high positive SHAP values (i.e. large impact on the prediction of increased mortality risk).

We also gain insights and witness the asymmetric effects that certain features and their values exert in influencing the final predicted risk scores. For instance, 'rx_risk_score', 'poor_eating_or_lack_of_appetite' and the 'pressure_ulcer_risk_score', exhibit a much stronger effect on elevating the predicted risk scores as their feature values increase, while the reverse effect is smaller on reducing risk as their feature values decrease. This is also in line with expectations, since for example, evidence of poor eating or a lack of appetite ought to have a greater effect on the model than a lack of evidence thereof. Other notable features are 'specific_health_conditions' and 'cognitive_performance_scale_score' which are largely consistent in signalling that a deterioration (higher values) in these features tends to also increase the predicted risk. However, a less coherent signal accompanies the 'depression_rating_scale_score', 'skin_integrity_score', 'faecal_incontinence' and 'falls_history' (investigated further below in dependence plots) with some signs of ambivalence with respect to predicted risk scores as the underlying

[6] While GBM was technically the top performing model, both it and XGB are essentially the same algorithm with slightly different implementations. The implementation of XGB, however, lends itself better for model transparency analysis given its integration with SHAP tools.

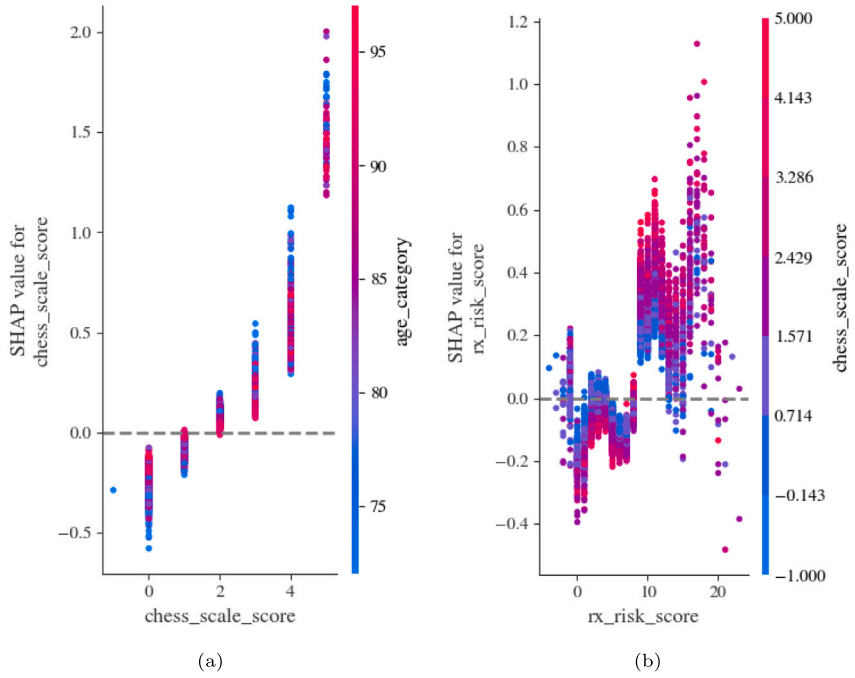


Fig. 2. Dependence plots showing pairwise feature interactions from the XGB model.

feature values change.

Two Dependence Plots are depicted in Fig. 2a-b, offering a deeper understanding of pairwise interactions between a selection of features and how this can be used to validate models in clinical settings. These visualizations depict how interactions of pairs of features influence the prediction of risk as their underlying values vary. As previously stated, these figures are based on the XGB model. For each of the selected features, the SHAP tool automatically selects the most interactive corresponding feature. The *x*-axis represents the values for a chosen feature, while the gradient colour bar on the *y*-axis represents the values of the counterpart feature. The interaction of both is depicted with respect to the magnitude of the impact they exert on the final prediction. The dashed horizontal line represents a neutral effect on the model output. Points above this line indicate an increase in mortality risk, while the opposite holds for values below the dashed line. The relative distance from the dashed line indicates the magnitude of the effect exerted on the mortality risk.

With an increasing 'chess_scale_score' in Fig. 2a, the mortality risk gradually increases. For lower values of the 'chess_scale_score', the interaction with increasing patient age tends to elevate overall risk. This relationship, however, does not seem to hold for higher values of the 'chess_scale_score'. A distinct pattern emerges for the 'rx_risk_score' feature in Fig. 2b as its values increase. A score of less than 10 for 'rx_risk_score' does not show a tendency to increase mortality risk. The interaction of lower values for this feature with increasing values of frailty represented by the 'chess_scale_score' tends not to elevate risk. However, an inflexion point occurs from 10 onwards for the 'rx_risk_score', at which point increasing values for both this feature and the 'chess_scale_score', interact to significantly elevate the mortality risk.

Clinical usage and application

Here, we transition from theoretical high-level modelling and inspection to an clinical decision making example, using two hypothetical patients as exemplars. The survival probability curve derived from the uncalibrated model and shown in Fig. 3, illustrates each patient's predicted survival trajectory in comparison to the cohort average. The figure indicates that the survival probabilities for patient B are significantly lower than those of patient A, and well below the cohort average across the entire timeframe of potential observation. This initial output allows clinicians to gauge individual patient risk in the context of broader population trends. However, while the initial risk-profile outputs are useful, additional insights are needed to unpack how and why the model is arriving at different risk predictions for specific patients.

Subsequently, the utilization of SHAP waterfall plots seen in Fig. 4a–b offers patient-level model interpretability. These plots reveal what the key features and their values are and how they influence the model's survival predictions for each patient. The emphasis here is on practicality: enabling clinicians to precisely comprehend the underpinnings of the model's output, ensuring that its insights can be validated, trusted and ultimately integrated into tailored patient management strategies. Through this approach, we demonstrate the confluence of advanced analytical tools with clinical utility, underscoring their role in optimizing patient care. These plots are best interpreted from bottom-up. The *y*-axis shows the most impactful features with their values, and their relative contributions. The starting point on the *x*-axis is the average or the expected risk for the whole cohort. Each feature pushes the risk to the left (to lower risk) or to the right (to increase risk) until all contributions are summed at the top row. For patient A in Fig. 4a, we can see that the patient's result on the use of mobility equipment increases their risk; however, their overall risk is significantly lowered by their independent mobilization, low pressure ulcer risk score, low number of ongoing health conditions and their female gender. Meanwhile, for patient B in Fig. 4b, it can be observed that their high risk is predominately driven by their high prescription risk score, limited mobility as well as their pressure ulcer risk score.

3.4. Calibrated time-specific survival models

The calibration plot with 95% confidence intervals for the target 6-month period (seen in Fig. 5) yields insights into the model's predictive accuracy. The plot features a calibration curve approximating the ideal 45-degree line, a sign of near-optimal calibration, where overall, the depicted model exhibits effective calibration from low to mid-range probabilities, while increased uncertainties around higher predicted

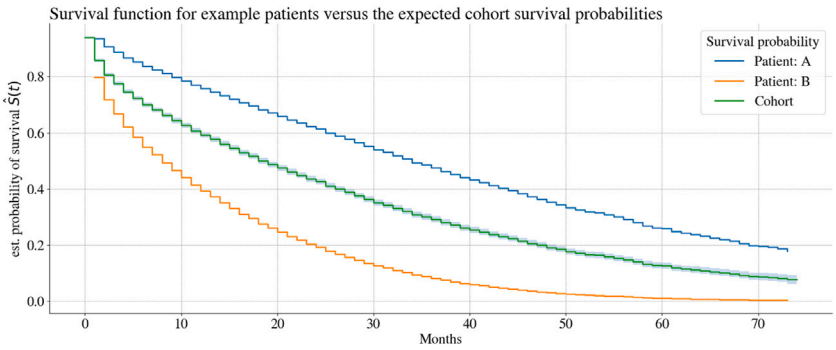
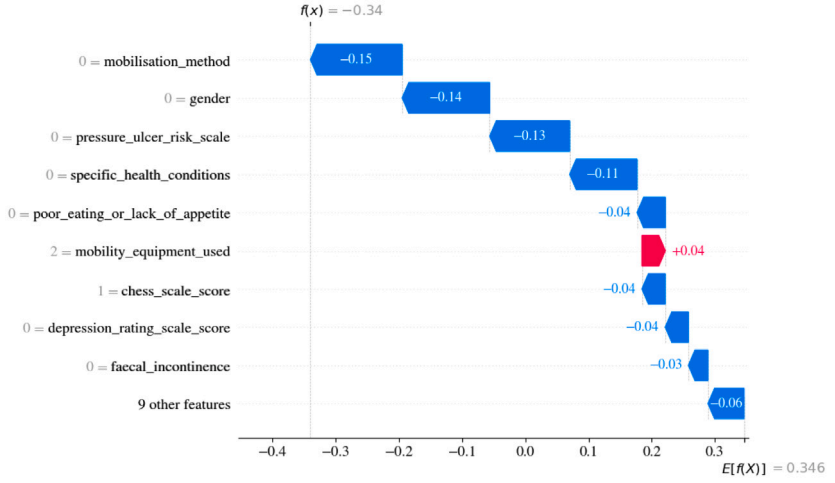
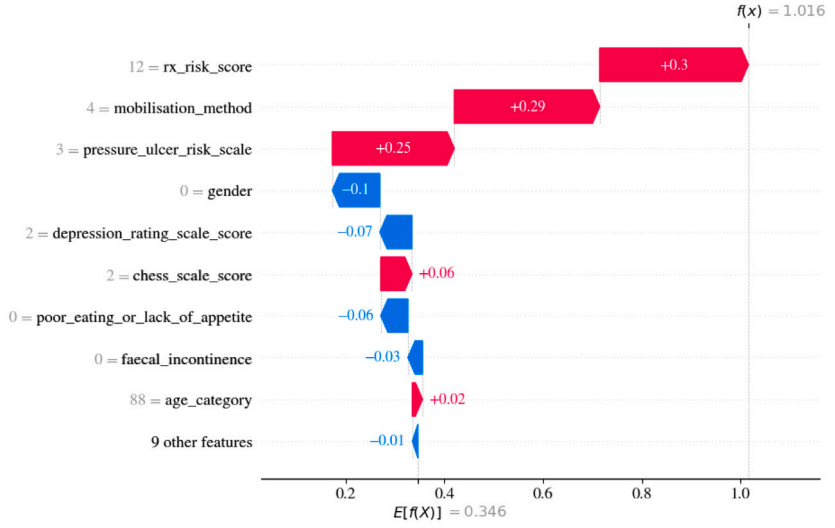


Fig. 3. Patient survival function from the GBM model depicting the risk of two exemplars versus the cohort.



(a) Patient A



(b) Patient B

Fig. 4. Waterfall plots from the XGB model showing explainability for two hypothetical patient examples.

survival probabilities (> 0.8) can also be seen. Beneath the calibration curve lies a histogram depicting the distribution of the predicted probabilities. The histogram intimates the density of predictions at different probability ranges. The shape of the distribution resembles a normal distribution with an albeit more pronounced tail for the lower probabilities, and a mean centred ~0.4. Additionally, the calibrated model was assessed by the Hosmer–Lemeshow statistic test which yielded a chi-square value of 12.6 and a *p*-value of 0.128, indicating satisfactory model calibration, and a reasonable congruence between the predicted probabilities and the observed outcomes. The implication is that the model neither significantly underfits nor overfits the data. Therefore, both the visual inspection and the statistical test are indicative of the model being effectively calibrated.

The performance metrics presented in Table 5 offer a multi-perspective evaluation of GBM models calibrated for survival analysis across varying temporal horizons. While the 6-month forecast is central

Table 5
Performance metrics of calibrated GBM models for time-specific forecasts with 95% CI reported in parentheses.

Forecast	Dynamic AUROC	IBS	C-index	Harrell C-index
1-month	0.794 (0.789–0.799)	0.296 (0.294–0.299)	0.715 (0.712–0.717)	0.674 (0.672–0.676)
3-month	0.765 (0.762–0.768)	0.280 (0.280–0.281)	0.717 (0.716–0.719)	0.676 (0.674–0.678)
6-month	0.746 (0.744–0.749)	0.259 (0.258–0.261)	0.716 (0.714–0.718)	0.675 (0.673–0.677)
12-month	0.726 (0.723–0.729)	0.239 (0.238–0.241)	0.720 (0.718–0.722)	0.680 (0.677–0.682)

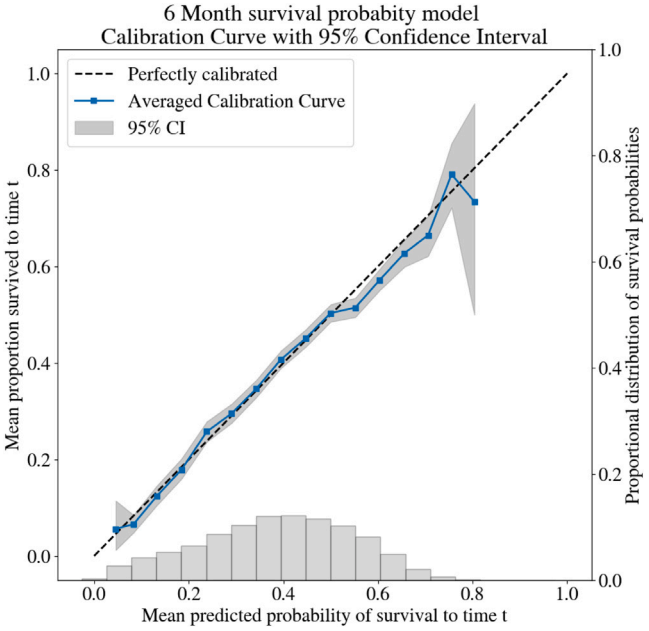


Fig. 5. 6-month GBM model calibration plot contrasting a perfectly calibrated pattern (dashed line) with the actual model, together with the distribution of the predicted probabilities.

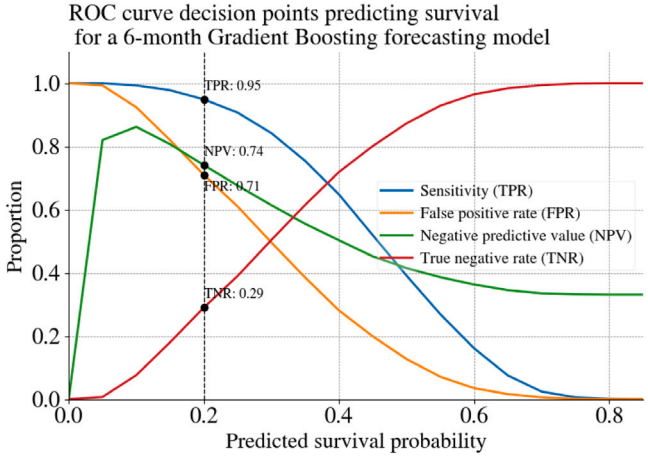


Fig. 6. ROC curve of calibrated GBM 6-month model predicting survival probabilities with 0.2 as an example threshold.

to this study, for completeness, the accuracies of 1-, 3- and 12-month calibrated models are also shown. The Dynamic AUROC serves as an emblematic metric for assessing a model's discriminative capability. The observed downward trend in AUROC values as we move from short-term to longer-term forecasts is indicative of an interplay between model sensitivity and the inherent heterogeneity of patient trajectories over time. A declining AUROC is often attributed to a natural increase in stochasticity of long-term forecasts, as can be seen in the table when contrasting the 1- and 12-month forecast accuracies.

The IBS serves as a gauge for model calibration. Though the observed trend of declining AUROC values over increasing prediction horizons aligns with the prevailing literature, signalling a diminishing discriminative power for long-term forecasts, intriguingly the model's IBS values improve (decrease) concurrently. This is counterintuitive given the conventional wisdom that long-term forecasts usually suffer from poor calibration. This paradox can however be explained through the lens of the bias–variance tradeoff: the results suggest that as the models become better calibrated over time, their variance reduces, thereby increasing bias and consequently reducing the models' discriminative power. The C-index and Harrell's index are used for their robustness in quantifying a model's ability to correctly rank-order individual risks. The results indicate the models' stability with respect to this across all forecast horizons.

3.5. Operationalization for clinical decision-making

The ROC curve illustrated in Fig. 6[7] for the calibrated 6-month GBM model shows both the performance characteristics of the model as well as an operational tool clinicians can use to adjust the decision-making threshold for identifying high-risk patients of mortality within 6 months that is suitable for different clinical contexts. It illustrates the balance between the true negative rate (TNR or specificity) and the true positive rate (TPR or sensitivity) achieved at various thresholds. As an example, the figure shows the clinical implications of adopting a 0.2 survival probability threshold (x-axis), equating to an 80% risk of death within the 6-month timeframe and the associated misclassification tradeoffs.

- **NPV**: quantifies the model's accuracy in predicting mortality. It is defined as the ratio of true non-survivors correctly predicted, to the total number of individuals predicted as non-survivors. Formally, NPV is given by:

$$NPV = \frac{TN}{TN + FN}, \tag{8}$$

  where TN represents the true negatives, i.e., individuals correctly predicted to not survive, and FN denotes false negatives, i.e., individuals who survived despite being predicted not to. At a specified threshold of 0.2, our model achieves an NPV of 74%, indicating that 74% of the residents predicted not to survive beyond six months at this threshold, indeed did not.
- **TNR**: measures the model's precision in identifying non-survivors. Although 0.29 is low, this level of caution is appropriate in a situation where under-estimating the likelihood of death within six months carries potentially less clinical and ethical risk than an over-estimation of risk.
- **TPR**: quantifies the model's ability to identify actual survivors, with an example threshold of 0.2 model output probability yielding a 95% TPR. This means that 95% of patients who survive

[7] Performance metrics presented are derived from validation on a separate dataset, ensuring a robust appraisal of the model's predictive capabilities.

beyond six months are accurately predicted by the model to survive.
- **FPR**: reflects the proportion of non-survivors incorrectly predicted as survivors. For this threshold, with a FPR of 71%, the model erroneously predicts survival in 71% of cases where the patient does not survive six months.

4. Discussion

Our study represents a significant contribution in gerontological research, notably for predicting survival rates in the residential aged care settings via advanced machine learning and XAI. Our approach goes beyond the current traditional methods, predominantly based on multivariate prognostic indices, by leveraging sophisticated predictive models whose evaluation metrics indicated satisfactory discriminatory power and robust statistical stability. Although prognostic indices have a long-standing history, one drawback is that their computation necessitates in-person consultations [16]. Since these consultations are inherently resource-intensive and thus do not scale effectively for identifying individuals requiring palliative care across a larger population such as those in aged care facilities, automated solutions such as machine learning are more appropriate. Results from our machine learning models generally surpassed existing prognostic models in residential aged-care populations reported in literature [8,46,47], while offering superior insights, reasoning and explainability for individualized patient predictions. The calibrated GBM models have demonstrated clinically useful predictive accuracy, which can be adjusted at different decision points on the ROC curve, in conjunction with model transparency through the integration of SHAP tools. These advancements facilitate a deeper understanding of individual patient prognoses, thereby enabling healthcare professionals to engage in more meaningful and informed discussions regarding end-of-life care preferences with residents and their families. A recent systematic literature review [48] has confirmed that machine learning is becoming more embedded within aged-care contexts. The study identified 70 studies conducted over a span of 12 years, exploring the application of machine learning techniques for the classification of various age-related health outcomes in older populations, including mortality. Neurodegenerative diseases were the primary focus in over 30% of studies, followed by mental health and the tracking of cardiovascular diseases, with both being the focus of 10% of studies each. However, notably, despite the extensive reliance on machine learning methods, only two studies employed XAI tools.

4.1. Ethical implications of AI use in palliative care decisions

Ethical considerations are central to the application of predictive AI-driven technologies in palliative care, particularly in ensuring fairness, accountability and respect for patient autonomy. XAI tools, such as SHAP used in this study, address these concerns by bringing to light the contribution of individual factors, such as comorbidities or cognitive impairment, to model predictions. The lucidity these tools offer mitigates biases inherent in machine learning models, fostering equitable and trustworthy decision-making. Medical literature is replete with examples where opaque AI systems, even when well-intentioned, perpetuate inequalities in healthcare access and outcomes [49,50], therefore, traceability characteristics of AI systems represent a core ethical requirement rather than simply a mere curious technological feature.

While it is relatively recent, XAI is not an unknown and untested technology, already having been successfully used in various fields in healthcare. [51] utilized XAI with machine learning on a large colorectal cancer registry, thereby identifying the ASA score and comorbidities such as COPD, asthma, hypertension, and myocardial infarction as key predictors of postoperative mortality which outperformed traditional scores. Meanwhile, [52] also reported that risk predictions from machine learning models for dementia and cognitive impairment are supported by the XAI tools in identifying variables that contribute to the prediction outcome.

In the palliative care context, by providing clinicians with explainable risk probabilities and insights, XAI enhances patient autonomy, enabling patients and families to make informed decisions aligned with their care preferences. This is an essential component, as in the context of palliative care, insights from these AI systems are not only for the care provider but also for patients and their families, allowing for truly shared decision-making. From the perspective of care providers, in the field where patient data and mortality risk prediction are complex, XAI provides a pathway to identify and address algorithmic bias at the level of individual cases. By exposing decision pathways, XAI helps ensure that predictions are non-discriminatory and clinically meaningful, and by instilling trust, these tools prove to be critical for a field where empathy and understanding are essential.

Collectively, both the findings and the suite of AI technologies used in this study, point to the potential to transform decision-making surrounding palliative care within aged care facilities, advocating for their role as primary providers of palliative services. The real-world applicability and impact of this study on the clinical adoption of these advanced tools, hold the promise of meaningfully improving decision-making processes in end-of-life care. Therefore, we believe that by integrating the workflow comprised of the set of technologies used in this study, a gap is bridged between theoretical models and their real-world application in end-of-life care, ensuring that decisions are transparent and aligned with patient preferences for delivering a higher level of care.

4.2. Future research

Subsequent studies should incorporate dynamic, time-varying features (such as new falls) to bolster both accuracy and responsiveness to changes. Crucially, validating these models across varied datasets and through clinical trials is essential to establish their generalizability. Research should also explore integrating novel features. Future work should also perform clinical trials using these patient-focused clinical decision tools with the aim of quantifying effects on the quality of end-of-life care, ensuring they are safe, effective, and aligned with patient needs.

4.3. Study limitations

The dataset contained missing data, which, despite sophisticated imputation, may cast some uncertainty on certain features and their interpretability, especially in SHAP analyses. The sole reliance on cross-sectional data at admission overlooks dynamic health changes which are potentially more indicative of mortality risks than static admission data. The patient cohort's homogeneity, sourced from a single provider, may to some degree limit the model's applicability in varied clinical settings. Furthermore, the study's technical validation requires an exploration into a real-world clinical integration which can only be satisfied in subsequent pragmatic clinical trials and user-centred designs to ensure the model's practical utility and effective incorporation into clinical workflows.

5. Conclusion

This study is the first comprehensive undertaking into developing machine learning models to predict survival probabilities for a general population of adults in residential aged care facilities, in combination with interpretable AI tools. This work conducted extensive experiments using numerous algorithms on a large dataset as well as a unique set of features, demonstrating the feasibility of developing robust predictive survival models in this setting which can be used by clinicians

for decision-making around appropriately targeted palliative care options. The use of advanced eXplainable AI (XAI) techniques was also uniquely integrated, showing how interpretability of the models and the explainability of their outputs can be realized to render machine learning suitable for clinical decision-making, where trust in the AI-driven prognostic tools is provided. Predictive models were calibrated for multiple time horizons, with an emphasis placed on the six-month survival probabilities post-admission to a residential aged care facility. TRIPOD reporting was adhered to and both the model parameters and code have been made publicly available. The proposed predictive framework, comprising the models and AI tools represents a significant step forward, offering a comprehensive approach to AI-driven healthcare for survival analysis in residential aged-care contexts and beyond.

CRediT authorship contribution statement

Teo Susnjak: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elise Griffin:** Writing – review & editing, Conceptualization.

Ethics statement

Patients and the public were not directly involved in this study. Ethics approval for this study was granted by the Aotearoa Research Ethics Committee (formerly New Zealand Ethics Committee, NZEC22_11) and noted by the Human Ethics Committee (Ohu Matatika 2) of Massey University.

Funding

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

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly, Overleaf's Writefull and ChatGPT occasionally to correct typographical and grammatical errors as well as to occasionally paraphrase existing text. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. All authors have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

The authors wish to express sincere thanks to Dr. Kathleen Potter for her crucial role in the inception, data provision, domain expertise and medical guidance of this research, a contribution that was pivotal to its success. The authors also thank Mitchell McCutcheon for his valuable feedback regarding previous modelling, data use and the integrity of different features in the dataset.

References

[1] J.C. Weeks, E.F. Cook, S.J. O'Day, L.M. Peterson, N. Wenger, D. Reding, F.E. Harrell, P. Kussin, N.V. Dawson, A.F. Connors, Jr., et al., Relationship between cancer patients' predictions of prognosis and their treatment preferences, Jama 279 (21) (1998) 1709–1714.

[2] N.A. Christakis, J.L. Smith, C.M. Parkes, E.B. Lamont, Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort studycommentary: Why do doctors overestimate? Commentary: Prognoses should be based on proved indices not intuition, Bmj 320 (7233) (2000) 469–473.

[3] A.A. Wright, B. Zhang, A. Ray, J.W. Mack, E. Trice, T. Balboni, S.L. Mitchell, V.A. Jackson, S.D. Block, P.K. Maciejewski, et al., Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment, Jama 300 (14) (2008) 1665–1673.

[4] M. Luppa, T. Luck, S. Weyerer, H.-H. König, E. Brähler, S.G. Riedel-Heller, Prediction of institutionalization in the elderly. A systematic review, Age Ageing 39 (1) (2010) 31–38.

[5] M. Omori, J. Jayasuriya, S. Scherer, B. Dow, M. Vaughan, S. Savvas, The language of dying: Communication about end-of-life in residential aged care, Death Stud. 46 (3) (2022) 684–694.

[6] A. Kelly, J. Conell-Price, K. Covinsky, I.S. Cenzer, A. Chang, W.J. Boscardin, A.K. Smith, Length of stay for older adults residing in nursing homes at the end of life, J. Am. Geriatr. Soc. 58 (9) (2010) 1701–1706.

[7] T. Sharp, E. Moran, I. Kuhn, S. Barclay, Do the elderly have a voice? Advance care planning discussions with frail and older individuals: a systematic literature review and narrative synthesis, Br. J. Gen. Pract. 63 (615) (2013) e657–e668.

[8] S. Zhang, K. Zhang, Y. Chen, C. Wu, Prediction models of all-cause mortality among older adults in nursing home setting: A systematic review and meta-analysis, Health Sci. Rep. 6 (6) (2023) e1309.

[9] J.M. Flacker, D.K. Kiely, A practical approach to identifying mortality-related factors in established long-term care residents, J. Am. Geriatr. Soc. 46 (8) (1998) 1012–1015, http://dx.doi.org/10.1111/j.1532-5415.1998.tb02759.x, arXiv:https://agsjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1532-5415.1998.tb02759.x, URL https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.1998.tb02759.x.

[10] J.M. Flacker, D.K. Kiely, Mortality-related factors and 1-year survival in nursing home residents, J. Am. Geriatr. Soc. 51 (2) (2003) 213–221.

[11] D. Porock, D. Parker Oliver, S. Zweig, M. Rantz, D. Mehr, R. Madsen, G. Petroski, Predicting death in the nursing home: development and validation of the 6-month minimum data set mortality risk index, J. Gerontol. Ser. A: Biol. Sci. Med. Sci. 60 (4) (2005) 491–498.

[12] D. Porock, D. Parker-Oliver, G.F. Petroski, M. Rantz, The MDS mortality risk index: The evolution of a method for predicting 6-month mortality in nursing home residents, BMC Res. Notes 3 (2010) 1–8.

[13] J.D. Niznik, S. Zhang, M.K. Mor, X. Zhao, M. Ersek, S.L. Aspinall, W.F. Gellad, J.M. Thorpe, J.T. Hanlon, L.J. Schleiden, et al., Adaptation and initial validation of minimum data set (MDS) mortality risk index to MDS version 3.0, J. Am. Geriatr. Soc. 66 (12) (2018) 2353–2359.

[14] R.T. Olender, S. Roy, P.S. Nishtala, Application of machine learning approaches in predicting clinical outcomes in older adults–a systematic review and meta-analysis, BMC Geriatr. 23 (1) (2023) 561.

[15] J. Allgaier, L. Mulansky, R.L. Draelos, R. Pryss, How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare, Artif. Intell. Med. 143 (2023) 102616, http://dx.doi.org/10.1016/j.artmed.2023.102616, URL https://www.sciencedirect.com/science/article/pii/S0933365723001306.

[16] E. Vu, N. Steinmann, C. Schröder, R. Förster, D.M. Aebersold, S. Eychmüller, N. Cihoric, C. Hertler, P. Windisch, D.R. Zwahlen, Applications of machine learning in palliative care: A systematic review, Cancers 15 (5) (2023) URL https://www.mdpi.com/2072-6694/15/5/1596.

[17] L. Wang, L. Sha, J.R. Lakin, J. Bynum, D.W. Bates, P. Hong, L. Zhou, Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions, JAMA Netw. Open 2 (7) (2019) e196972–e196972.

[18] S. Mostafaei, M.T. Hoang, P.G. Jurado, H. Xu, L. Zacarias-Pons, M. Eriksdotter, S. Chatterjee, S. Garcia-Ptacek, Machine learning algorithms for identifying predictive variables of mortality risk following dementia diagnosis: a longitudinal cohort study, Sci. Rep. 13 (1) (2023) 9480.

[19] I. Maouche, L.S. Terrissa, K. Benmohammed, N. Zerhouni, An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data, IEEE Trans. Biomed. Eng. 70 (12) (2023) 3321–3329, http://dx.doi.org/10.1109/TBME.2023.3282840.

[20] F. Berloco, P.M. Marvulli, V. Suglia, S. Colucci, G. Pagano, L. Palazzo, M. Aliani, G. Castellana, P. Guido, G. D'Addio, V. Bevilacqua, Enhancing survival analysis model selection through XAI(t) in healthcare, Appl. Sci. 14 (14) (2024) http://dx.doi.org/10.3390/app14146084, URL https://www.mdpi.com/2076-3417/14/14/6084.

[21] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement, Circulation 131 (2) (2015) 211–219.

[22] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, Association for Computing Machinery, New York, NY, USA, 2005, pp. 625–632, http://dx.doi.org/10.1145/1102351.1102430.

[23] S. Rajaraman, P. Ganesan, S. Antani, Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks, PLoS One 17 (1) (2022) 1–23, http://dx.doi.org/10.1371/journal.pone.0262838.

[24] N.L. Pratt, M. Kerr, J.D. Barratt, A. Kemp-Casey, L.M.K. Ellett, E. Ramsay, E.E. Roughead, The validity of the rx-risk comorbidity index using medicines mapped to the anatomical therapeutic chemical (ATC) classification system, BMJ Open 8 (4) (2018) e021122.

[25] InterRAI Long-Term Care Facilities (LTCF) Assessment Form and User's Manual, Australian Edition, 9.1.2, InterRAI, 2020.

[26] J.P. Hirdes, J.W. Poss, L. Mitchell, L. Korngut, G. Heckman, Use of the interRAI CHESS scale to predict mortality among persons with neurological conditions in three care settings, PLoS One 9 (6) (2014) e99066.

[27] J.A. Ogarek, E.M. McCreedy, K.S. Thomas, J.M. Teno, P.L. Gozalo, Minimum data set changes in health, end-stage disease and symptoms and signs scale: a revised measure to predict mortality in nursing home residents, J. Am. Geriatr. Soc. 66 (5) (2018) 976–981.

[28] C. Travers, G. Byrne, N. Pachana, K. Klein, L. Gray, Validation of the interRAI cognitive performance scale against independent clinical diagnosis and the mini-mental state examination in older hospitalized patients, J. Nutr. Health Aging 17 (2013) 435–439.

[29] K. Penny, A. Barron, A.-M. Higgins, S. Gee, M. Croucher, G. Cheung, Convergent validity, concurrent validity, and diagnostic accuracy of the interRAI depression rating scale, J. Geriatr. Psychiatry Neurol. 29 (6) (2016) 361–368.

[30] J. Poss, K.M. Murphy, M.G. Woodbury, H. Orsted, K. Stevenson, G. Williams, S. MacAlpine, N. Curtin-Telegdi, J.P. Hirdes, Development of the interRAI pressure ulcer risk scale (PURS) for use in long-term care and home care settings, BMC Geriatr. 10 (2010) 1–10.

[31] N. Shah, C. Konchak, D. Chertok, L. Au, A. Kozlov, U. Ravichandran, P. McNulty, L. Liao, K. Steele, M. Kharasch, C. Boyle, T. Hensing, D. Lovinger, J. Birnberg, A. Solomonides, L. Halasyamani, Clinical analytics prediction engine (CAPE): Development, electronic health record integration and prospective validation of hospital mortality, 180-day mortality and 30-day readmission risk prediction models, PLoS One 15 (8) (2020) 1–15, http://dx.doi.org/10.1371/journal.pone.0238065.

[32] S. Van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification, Stat. Methods Med. Res. 16 (3) (2007) 219–242.

[33] M. Mera-Gaona, U. Neumann, R. Vargas-Canas, D.M. López, Evaluating the impact of multivariate imputation by MICE in feature selection, PLoS One 16 (7) (2021) e0254720.

[34] C.M. Caruso, V. Guarrasi, S. Ramella, P. Soda, A deep learning approach for overall survival prediction in lung cancer with missing values, Comput. Methods Programs Biomed. 254 (2024) 108308, http://dx.doi.org/10.1016/j.cmpb.2024.108308, URL https://www.sciencedirect.com/science/article/pii/S016926072400302X.

[35] S. Pölsterl, Scikit-survival: A library for time-to-event analysis built on top of scikit-learn, J. Mach. Learn. Res. 21 (212) (2020) 1–6, URL http://jmlr.org/papers/v21/20-729.html.

[36] XGBoost Development Team, XGBoost: Extreme gradient boosting, https://pypi.org/project/xgboost/, Python Package.

[37] J. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. (2001).

[38] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[39] F.E. Harrell, K.L. Lee, D.B. Mark, Evaluating the yield of medical tests, JAMA (1982).

[40] F.E. Harrell, Regression Modeling Strategies, Springer, 2015.

[41] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology (1982).

[42] E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Multistate survival models for panel data: The msm package for R, J. Stat. Softw. (1999).

[43] C.R. Manz, J. Chen, M. Liu, C. Chivers, S.H. Regli, J. Braun, M. Draugelis, C.W. Hanson, L.N. Shulman, L.M. Schuchter, N. O'Connor, J.E. Bekelman, M.S. Patel, R.B. Parikh, Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer, JAMA Oncol. 6 (11) (2020) 1723–1730, http://dx.doi.org/10.1001/jamaoncol.2020.4331, arXiv:https://jamanetwork.com/journals/jamaoncology/articlepdf/2770698/jamaoncology_manz_2020_oi_200068_1604947261.21354.pdf.

[44] D.W. Hosmer, Jr., S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, third ed., John Wiley & Sons, 2013.

[45] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st Int. Conf. on Neural Information Processing Systems, NIPS '17, Curran Assoc., 2017, pp. 4768–4777.

[46] R.L. Kruse, D. Parker Oliver, D.R. Mehr, G.F. Petroski, D.L. Swenson, S.C. Zweig, Using mortality risk scores for long-term prognosis of nursing home residents: caution is recommended, J. Gerontol. Ser. A: Biomed. Sci. Med. Sci. 65 (11) (2010) 1235–1241.

[47] J.T. van der Steen, S.L. Mitchell, D.H. Frijters, R.L. Kruse, M.W. Ribbe, Prediction of 6-month mortality in nursing home residents with advanced dementia: validity of a risk score, J. Am. Med. Dir. Assoc. 8 (7) (2007) 464–468.

[48] A. Das, P. Dhillon, Application of machine learning in measurement of ageing and geriatric diseases: a systematic review, BMC Geriatr. 23 (1) (2023) 841.

[49] H. Siala, Y. Wang, SHIFTing artificial intelligence to be responsible in healthcare: A systematic review, Soc. Sci. Med. 296 (2022) 114782, http://dx.doi.org/10.1016/j.socscimed.2022.114782, URL https://www.sciencedirect.com/science/article/pii/S0277953622000855.

[50] R.J. Chen, J.J. Wang, D.F. Williamson, T.Y. Chen, J. Lipkova, M.Y. Lu, S. Sahai, F. Mahmood, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nat. Biomed. Eng. 7 (6) (2023) 719–742.

[51] T. van den Bosch, A.-L.K. Warps, M.P.M. de Nerée tot Babberich, C. Stamm, B.F. Geerts, L. Vermeulen, M.W.J.M. Wouters, J.W.T. Dekker, R.A.E.M. Tollenaar, P.J. Tanis, D.M. Miedema, D.C. Audit, Predictors of 30-day mortality among dutch patients undergoing colorectal cancer surgery, 2011–2016, JAMA Netw. Open 4 (4) (2021) http://dx.doi.org/10.1001/jamanetworkopen.2021.7737, e217737–e217737.

[52] W.Y. Tan, C. Hargreaves, C. Chen, S. Hilal, A machine learning approach for early diagnosis of cognitive impairment using population-based data, J. Alzheimer's Dis. 91 (1) (2023) 449–461.