PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Naïve Bayes is an interpretable and predictive machine learning algorithm in predicting osteoporotic hip fracture in-hospital mortality compared to other machine learning algorithms

Jo-Wai Douglas Wang[1,2]*

1 Department of Geriatric Medicine, The Canberra Hospital, ACT Health, Canberra, Australia, **2** The Australian National University Medical School, Canberra, Australia

* jo-waidouglas.wang@anu.edu.au

OPEN ACCESS

Citation: Wang J-WD (2025) Naïve Bayes is an interpretable and predictive machine learning algorithm in predicting osteoporotic hip fracture in-hospital mortality compared to other machine learning algorithms. PLOS Digit Health 4(1): e0000529. https://doi.org/10.1371/journal.pdig.0000529

Editor: Mathew V. Kiang, Stanford University School of Medicine, UNITED STATES OF AMERICA

Received: May 9, 2024

Accepted: November 8, 2024

Published: January 2, 2025

Data Availability Statement: The data used for analysis in this study is available to researchers by written request to the approving Australian Capital Territory Human Research Ethics Committee at ethics@act.gov.au (ethics reference number: 2023.LRE.00063). Unfortunately, data cannot be made available on a public repository due to the conditions of approval set by the approving ethics committee as it contains potentially identifying and sensitive patient information.

Abstract

Osteoporotic hip fractures (HFs) in the elderly are a pertinent issue in healthcare, particularly in developed countries such as Australia. Estimating prognosis following admission remains a key challenge. Current predictive tools require numerous patient input features including those unavailable early in admission. Moreover, attempts to explain machine learning [ML]-based predictions are lacking. Seven ML prognostication models were developed to predict in-hospital mortality following minimal trauma HF in those aged $\geq$ 65 years of age, requiring only sociodemographic and comorbidity data as input. Hyperparameter tuning was performed via fractional factorial design of experiments combined with grid search; models were evaluated with 5-fold cross-validation and area under the receiver operating characteristic curve (AUROC). For explainability, ML models were directly interpreted as well as analysed with SHAP values. Top performing models were random forests, naïve Bayes [NB], extreme gradient boosting, and logistic regression (AUROCs ranging 0.682–0.696, p>0.05). Interpretation of models found the most important features were chronic kidney disease, cardiovascular comorbidities and markers of bone metabolism; NB also offers direct intuitive interpretation. Overall, NB has much potential as an algorithm, due to its simplicity and interpretability whilst maintaining competitive predictive performance.

Author summary

Osteoporotic hip fractures are a critical health issue in developed countries. Preventative measures have ameliorated this issue somewhat, but the problem is expected to remain in main due to the aging population. Moreover, the mortality rate of patients in-hospital remains unacceptably high, with estimates ranging from 5–10%. Thus, a risk stratification tool would play a critical role in optimizing care by facilitating the identification of the

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

susceptible elderly in the community for prevention measures and the prioritisation of such patients early during their hospital admission. Unfortunately, such a tool has thus far remained elusive, despite forays into relatively exotic algorithms in machine learning. There are three major drawbacks (1) most tools all rely on information typically unavailable in the community and early during admission (for example, intra-operative data), limiting their potential use in practice, (2) few studies compare their trained models with other potential algorithms and (3) machine learning models are commonly cited as being 'black boxes' and uninterpretable. Here it is shown that a Naïve Bayes model, trained using only sociodemographic and comorbidity data of patients, performs on par with the more popular methods lauded in literature. The model is interpretable through direct analysis; the comorbidities of chronic kidney disease, cardiovascular, and bone metabolism were identified as being important features contributing to the likelihood of deaths. An algorithm-agnostic approach to machine learning model interpretation is also shown. This study shows the potential for Naïve Bayes in predicting elderly patients at risk of death during an admission for hip fracture.

1. Introduction

The osteoporotic hip fracture (HF) is a global issue with an estimated financial burden of 17 billion USD for the United States in 2002 and projected burden of £3.62 in 2023 for the United Kingdom [1,2]. Estimates for short-term (in-hospital) mortality following HF have been placed in the vicinity of 2–10%, with an estimated mortality rate of 2.7% for HF hospitalisation in Australia [3–5]. In developed countries, though preventative measures (targeting reduction of hip fracture risk factors such as osteoporosis and falls) have reduced the age-standardized incidence rate of hip fractures, the absolute rate is increasing due to the ageing population [6]. In Australia, for instance, hospitalisations for HF in the elderly increased by almost 20% between 2006–07 and 2015–16 from 15 900 to 18 700 respectively [5]. With the trend towards an aged population expected to continue, including in Australia, HFs in the elderly will remain a relevant, and increasingly pressing challenge in healthcare.

One key aspect in the management of HF is the prognostication of poor short-term outcomes. There exists a substantial amount of analysis from traditional statistical methods (such as logistical regression, LR) in identifying key risk factors for predicting poor outcomes, notably mortality, following HF and scoring tools that have risen to prominence are the Nottingham Hip Fracture Score (NHFS) and the orthopaedic- Physiological and Operative Severity Score for the enUmeration of Mortality and Morbidity (O-POSSUM) [7–10]. Most of these tools require a combination of both clinical, laboratory and intra-operative data; and the lack of laboratory and intra-operative data early during admission limits the use of such tools in early risk stratification.

Non-traditional mathematical algorithms, especially those associated with artificial intelligence (AI) and machine learning (ML), have become increasingly utilized in healthcare. A variety of ML algorithms, including regression-based methods, decision-tree based methods (i.e. decision trees [DT], Random Forests [RF], eXtreme Gradient Boosting [XGB] implementation), neural networks (NN), Naïve Bayes and support vector machines (SVM) have been used in the prognostication of patients in the general peri-operative [11–16] and peri-HF [17–20] period with varying degrees of success. However, most of these tools require data that are not readily available on admission (such as intra-operative data and laboratory data), much

like the tools developed from traditional statistical methods, and most do not predict short-term in-hospital mortality following HF.

Moreover, there is a scarcity of studies that developed multiple different machine learning algorithms in HF prediction, and compared the end-results with one another. No study has trained and compared multiple machine learning algorithms for prediction of short-term outcomes (e.g. in-hospital mortality) following HFs. Indeed, even in task of long-term prognostication, only one study attempted to train and compare different classes of algorithms (SVM, NB and LR) in the in their ability to predict 1-year mortality post-HF [17]. Tree-based methods have received the majority of attention. An algorithm that has remarkable potential is Naïve Bayes, which is based on Bayes theorem, with the additional 'naïve' assumption that features are conditionally independent. It has been applied successfully across a wide variety of tasks in natural language processing (e.g. detection of spam email [21], text sentiment analysis, text/document classification) as well as in the medical field (e.g. the prognostication in cirrhotic patients following transjugular intrahepatic portosystemic shunt [22], prediction of 30-day mortality following HF [23], prediction of osteonecrosis of femoral head with cannulated screw fixation [24] and prediction of mortality in post-surgical intensive care unit patients [25]).

While predictive ability is an important characteristic of any prognostic tool, it is increasingly recognized that a desirable attribute of machine learning algorithm is that they are interpretable (or 'explainable') especially as ML models become increasingly complex [26,27]. Recognition of this issue has led to the development of the subfield of 'interpretable' ML and, in particular, the development and application of the SHapley Additive exPlanations (SHAP), an approach based on cooperative game theory [28–34].

The goal, was to train multiple ML models, specifically Bernoulli Naïve Bayes (NB), DT, RF, XGB, SVM, logistic regression (LR) and the multi-layer perceptron (MLP, a 3-layer NN) to predict in-hospital mortality for the elderly admitted with HF. The focus of this study is on using only those patient features that are readily available in the early phases during a hospital admission, i.e. sociodemographic and comorbidity data. The performances of each model would be compared to identify the most predictive algorithm. Finally, each predictive tool would be analysed via direct interpretation of model and with calculation of SHAP values.

2. Results

2.1. Patient cohort characteristics

Of the 3625 patients in the cohort, age was distributed non-normally with median age of 84 (interquartile range of 10 years) and females comprising 2730 (75.3%); 189 (5.2%) had in-hospital mortality. The most common comorbidity was hypertension (HTN, at 2045 [56.4%]). Details are present in Table 1 (with abbreviations defined below).

2.2. Model performance–training

The model with the highest area under the receiver operating characteristic (AUROC) was MLP (AUROC 0.828) followed by LR, RF, XGB and NB (0.733, 0.730, 0.726 and 0.725 respectively, all $p>0.05$), then DT (AUROC of 0.697) and finally SVM (AUROC 0.533).

The model with greatest area under the precision-recall curve (AUPRC) was MLP (AUPRC 0.245), followed by LR, XGB and RF (AUPRCs of 0.134, 0.133 and 0.130 respectively, $p>0.05$), NB (AUPRC 0.124), DT (AUPRC of 0.094) and finally SVM (AUPRC of 0.058). Details are present in Table 2 and Table 3.

Table 1. Sociodemographic features, outcomes of HF cohort.

Variable	Total Cohort (N = 3625)	Female (N = 2730, 75.31%)	Male (N = 895, 24.69%)	p value[1]
Sociodemographic features				
Age (median [IQR])	84 [10]	85 [10]	82 [12]	<0.001
Aged > 80 years (n,%)	2457, 67.8%	1937, 71.0%	520, 58.1%	<0.001
PRCF resident (n,%)	1208, 33.3%	950, 34.8%	258, 28.8%	0.001
Smoker (n,%)	180, 5.0%	124, 4.5%	56, 6.3%	0.050
Alcohol overuse (n,%)[2]	144, 4.0%	60, 2.2%	84, 9.4%	<0.001
Walking aids user (n,%)	1300, 35.9%	999, 36.6%	301, 33.6%	0.116
Comorbidities features				
HTN, (n,%)	2045, 56.4%	1606, 58.8%	439, 49.1%	<0.001
Anaemia (n,%)	1531, 42.2%	1051, 38.5%	480, 53.6%	<0.001
CKD (n,%)	1444, 39.9%	1106, 40.5%	338, 37.8%	0.152
Dementia (n,%)	1117, 30.8%	858, 31.4%	259, 28.9%	0.172
CAD (n,%)	1073, 29.6%	750, 27.5%	323, 36.1%	<0.001
History of AMI (n,%)	287, 7.9%	191, 7.0%	96, 10.7%	<0.001
AF (n,%)	702, 19.4%	513, 18.8%	189, 21.1%	0.139
COPD (n,%)	561, 15.5%	385, 14.1%	176, 19.7%	<0.001
T2DM (n,%)	482, 13.3%	325, 11.9%	157, 17.5%	<0.001
OP (n,%)	478, 13.2%	410, 15.0%	68, 7.6%	<0.001
CVA (n,%)	431, 11.9%	323, 11.8%	108, 12.1%	0.897
TIA (n,%)	309, 8.5%	227, 8.3%	82, 9.2%	0.474
PD (n,%)	172, 4.7%	97, 3.6%	75, 8.4%	<0.001
Malignancy (n,%)	82, 2.3%	52, 1.9%	30, 3.4%	0.017
PTH>6.8pmol/L	1684, 46.5%	1275, 46.7%	409, 45.7%	0.628
25(OH)vitamin D≤25nmol/L	610, 16.8%	467, 17.1%	143, 16.0%	0.464
25(OH)vitamin D≤50nmol/L	1659, 45.8%	1235, 45.2%	424, 47.4%	0.283
Outcome				
Died (n,%)	189, 5.2%	130, 4.8%	59, 6.6%	0.040

[1]Pearson's Chi-squared test (Yates corrected).

[2]Use>3 times a week.

Abbreviations: PRCF, permanent residential care facility; HTN, hypertension; CKD, chronic kidney disease; CAD, coronary artery disease; AMI, acute myocardial infarction; AF, atrial fibrillation; COPD, chronic obstructive pulmonary disease; T2DM, type 2 diabetes mellitus; OP, osteoporosis; CVA, cerebrovascular accident; TIA, transient ischaemic attack; PD, Parkinson's disease; PTH, parathyroid hormone.

https://doi.org/10.1371/journal.pdig.0000529.t001

Table 2. Model performance (training phase).

	AUROC			AUPRC		
	Mean	**STD**	**95%CI**	**Mean**	**STD**	**95%CI**
SVM	0.533	0.029	0.475–0.591	0.058	0.004	0.050–0.067
NB	0.725	0.007	0.711–0.739	0.124	0.003	0.117–0.131
LR	0.733	0.008	0.717–0.750	0.134	0.004	0.127–0.141
DT	0.697	0.004	0.690–0.704	0.094	0.001	0.093–0.095
RF	0.730	0.007	0.716–0.745	0.130	0.003	0.125–0.136
XGB	0.726	0.007	0.711–0.741	0.133	0.005	0.122–0.144
MLP	0.828	0.008	0.813–0.844	0.245	0.030	0.186–0.305

https://doi.org/10.1371/journal.pdig.0000529.t002

Table 3. Comparison of model performance during training. (A)–AUROC (B)–AUPRC.

(A) AUROC

	t-test statistic							*t*-test *p*-value						
models	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**
SVM	-	-14.391	-14.866	-12.527	-14.766	-14.466	-21.927	-	0.000	0.000	0.000	0.000	0.000	0.000
NB	-	-	-1.683	7.766	-1.129	-0.226	-21.666	-	-	0.131	0.000	0.291	0.827	0.000
LR	-	-	-	9.000	0.631	1.472	-18.776	-	-	-	0.000	0.546	0.179	0.000
DT	-	-	-	-	-9.153	-8.043	-32.750	-	-	-	-	0.000	0.000	0.000
RF	-	-	-	-	-	0.904	-20.614	-	-	-	-	-	0.393	0.000
XGB	-	-	-	-	-	-	-21.456	-	-	-	-	-	-	0.000
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

(B) AUPRC

	t-test statistic							*t*-test *p*-value						
models	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**
SVM	-	-29.516	-30.042	-19.524	-40.249	-26.191	-13.816	-	0.000	0.000	0.000	0.000	0.000	0.000
NB	-	-	-4.472	21.213	-4.472	-3.451	-8.974	-	-	0.002	0.000	0.002	0.009	0.000
LR	-	-	-	21.693	2.236	0.349	-8.201	-	-	-	0.000	0.056	0.736	0.000
DT	-	-	-	-	-80.498	-17.103	-11.249	-	-	-	-	0.000	0.000	0.000
RF	-	-	-	-	-	-1.342	-8.572	-	-	-	-	-	0.217	0.000
XGB	-	-	-	-	-	-	-8.234	-	-	-	-	-	-	0.000
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

https://doi.org/10.1371/journal.pdig.0000529.t003

2.3. Model performance–test

The models with highest AUROC were RF, NB, XGB and LR (AUROCs of 0.696, 0.694, 0.689 and 0.682 respectively, all p>0.05;) followed by MLP and DT (AUROCs of 0.618, 0.616 respectively, p>0.05) and finally SVM (AUROC of 0.499).

The model(s) with highest AUPRC were NB, XGB, RF and LR (AUPRCs of 0.113, 0.113, 0.112, 0.112 respectively, p>0.05), MLP and DT (AUPRCs of 0.077, 0.070 respectively, p>0.05), and SVM (AUPRC of 0.054)–see Table 4, Table 5 and S1 Fig.

2.4. Feature importance–Model interpretation

Feature importance rankings (1 being the most important) according to each model can be found in Table C in S3 Appendix. Corresponding coefficients for NB, LR, XGB and RF can be found in Fig 1.

For the LR model, the 5 most important patient features in prediction of mortality were presence of CKD, vitamin D deficiency (≤25nmol/L), advanced age (>80 years), COPD, and

Table 4. Model performance during testing (5-fold cross-validation).

	AUROC			AUPRC—Average precision		
	Mean	**STD**	**95%CI**	**Mean**	**STD**	**95%CI**
SVM	0.499	0.048	0.404–0.594	0.054	0.006	0.041–0.067
NB	0.694	0.024	0.646–0.742	0.113	0.019	0.076–0.151
LR	0.682	0.034	0.614–0.750	0.112	0.019	0.074–0.150
DT	0.616	0.028	0.560–0.671	0.070	0.006	0.058–0.081
RF	0.696	0.030	0.636–0.756	0.112	0.024	0.064–0.161
XGB	0.689	0.025	0.640–0.738	0.113	0.024	0.065–0.160
MLP	0.618	0.027	0.565–0.671	0.077	0.007	0.063–0.090

https://doi.org/10.1371/journal.pdig.0000529.t004

Table 5. Comparison of model performance on testing. (A)–AUROC (B)–AUPRC.

(A) AUROC

	t-test statistic							*t*-test *p*-value					
models	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**
SVM	-	-8.125	-6.957	-4.708	-7.782	-7.850	-4.832	-	0.000	0.000	0.002	0.000	0.000	0.001
NB	-	-	0.645	4.729	-0.116	0.323	4.704	-	-	0.537	0.001	0.910	0.755	0.002
LR	-	-	-	3.351	-0.690	-0.371	3.296	-	-	-	0.010	0.509	0.720	0.011
DT	-	-	-	-	-4.359	-4.349	-0.115	-	-	-	-	0.002	0.002	0.911
RF	-	-	-	-	-	0.401	4.321	-	-	-	-	-	0.699	0.003
XGB	-	-	-	-	-	-	4.315	-	-	-	-	-	-	0.003
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

(B) AUPRC

	t-test statistic							*t*-test *p*-value					
models	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**	**SVM**	**NB**	**LR**	**DT**	**RF**	**XGB**	**MLP**
SVM	-	-6.621	-6.509	-4.216	-5.242	-5.333	-5.578	-	0.000	0.002	0.003	0.001	0.001	0.001
NB	-	-	0.083	4.826	0.073	0.000	3.976	-	-	0.936	0.001	0.944	1.000	0.004
LR	-	-	-	4.713	0.000	-0.073	3.865	-	-	-	0.002	1.000	0.944	0.005
DT	-	-	-	-	-3.796	-3.887	-1.698	-	-	-	-	0.005	0.005	0.128
RF	-	-	-	-	-	-0.066	3.130	-	-	-	-	-	0.949	0.014
XGB	-	-	-	-	-	-	3.220	-	-	-	-	-	-	0.012
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

https://doi.org/10.1371/journal.pdig.0000529.t005

AF. In the SVM model the 5 most important patients features in prediction of mortality were advanced age (>80 years), CKD, vitamin D insufficiency (≤50nmol/L), anaemia, and use of walking aids. For the NB model, the 5 most important features in mortality prediction were history of MI, AF, CKD and CAD. For the DT model the 5 most important features in mortality prediction were presence of CKD, hyperparathyroidism (PTH>6.8pmol/L), CAD, dementia and advanced age (>80 years). For the RF model the 5 most important features in mortality prediction were CKD, hyperparathyroidism (PTH>6.8pmol/L), CAD, dementia and advanced age (>80 years). For the XGB model the 5 most important features in mortality prediction were CKD, CAD, advanced age (>80 years), PTH>6.8pmol/L and AF. Finally, for the MLP model the 5 most important features in mortality prediction were AF, CKD, male sex, dementia, and MI.

2.5. Feature importance–SHAP analysis

Features were also ranked by the mean absolute SHAP values (Figs 2–7).

For the LR model, the 5 most predictive patient features for mortality in order from highest magnitude to lowest, based on mean SHAP values, were CKD, advanced age (>80 years), hyperparathyroidism (PTH>6.8pmol/L), CAD, and residency from PRCF. Absence of any of these features had a negative SHAP value (i.e. a negative contribution) on the model outcome (in-hospital mortality); the magnitude of this impact was consistent across all patients. Likewise, the presence of any of these features always had a positive SHAP value (i.e. an additive contribution) on in-hospital mortality. The magnitude of this effect was again consistent across all patients.

For the NB model, the 5 most predictive features were CKD, AF, MI, residency from PRCF, CAD. Again, absence of any of these features most commonly had a negative impact on in-hospital mortality; the magnitude of this effect varied among patients. The presence of any of the above 5 features had a positive contribution to the prediction of in-hospital mortality; similarly, the magnitude of this effect varied significantly among patients.

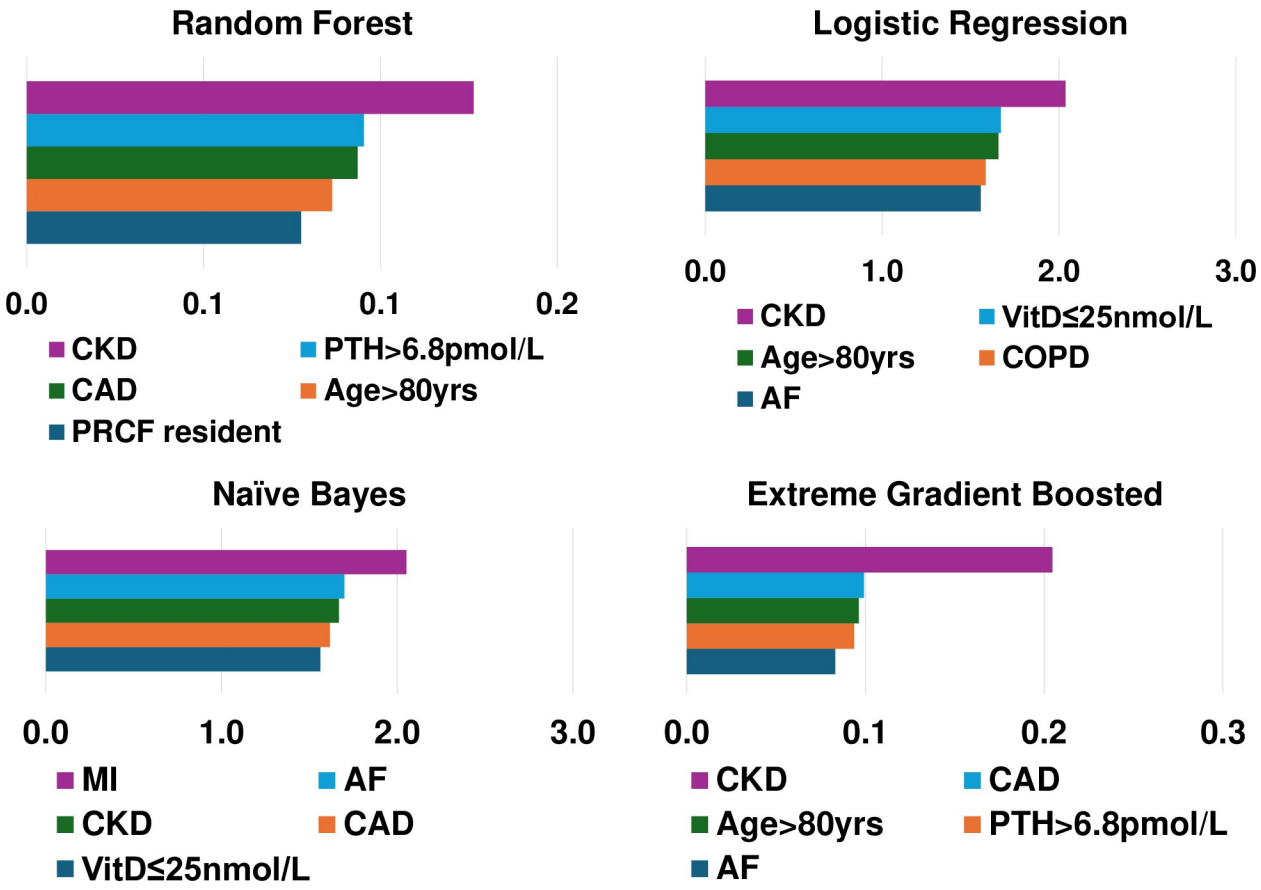

Fig 1. Feature importance based on model interpretation. The presence of CKD, cardiovascular comorbidities (either CAD or AF),deranged markers of bone metabolism (PTH>6.8pmol/Land vitamin D≤25nmol/L) and advanced age contributed the greatest amount to predictions across models.

https://doi.org/10.1371/journal.pdig.0000529.g001

For the DT model, the 5 most predictive features were CKD, hyperparathyroidism (PTH>6.8pmol/L), advanced age (>80 years), presence of CAD and vitamin D deficiency (≤25nmol/L). Presence of these five comorbidities had a positive contribution to prediction of in-hospital mortality and, conversely their absence had a negative contributory effect on prediction. Interestingly, absence of T2DM had an additive effect and presence of T2DM had a negative effect on mortality prediction. The magnitude of contributions that each of the 5 variables had varied among different patients. Finally, it is noteworthy that all other comorbidities had little to no influence on patient outcomes.

For the RF model, the 5 most predictive features were CKD, hyperparathyroidism (PTH>6.8pmol/L), CAD, advanced age (>80 years) and residence from PRCF. The presence of these features increased likelihood of mortality and conversely absence decreased the likelihood of mortality; there was only a minor variation of contribution from each feature for each patient.

For the XGB model, the 5 most predictive patient features were advanced age (>80 years), vitamin D deficiency (≤25nmol/L), CKD, CAD and hyperparathyroidism (PTH>6.8pmol/L). The presence (and absence) of any of these features increased (or decreased) the likelihood of mortality. For each feature, there was only mild variation in the magnitude of contributions among patients.

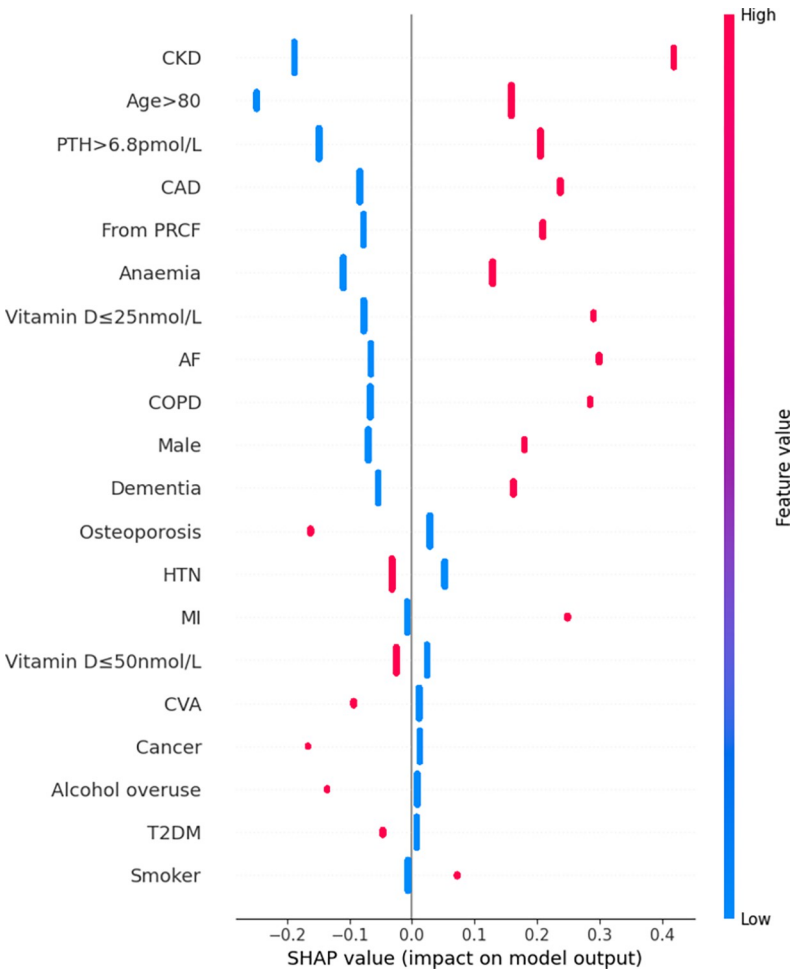

Fig 2. Summary plot SHAP values for patient comorbidities–Logistic Regression. Each point on the plot represents a SHAP value for an individual patient's comorbidity (SHAP value is on x-axis, corresponding comorbidity on the y-axis. Positive SHAP values corresponds to a positive/additive contribution to the prediction (i.e. in-hospital mortality); conversely a negative SHAP value corresponds to a negative/subtractive contribution. Colours of points represents feature values: magenta/red corresponded to a value of '1' (i.e. presence of the comorbidity) and blue corresponding to value '0' (i.e. absence of comorbidity).

https://doi.org/10.1371/journal.pdig.0000529.g002

Finally, from the MLP model, the 5 most predictive patient features were male sex, advanced age (>80 years), CVA, HTN and TIA. The presence (or, conversely, the absence) of any of these features except for HTN were associated with an increased (decreased) likelihood of mortality; presence (or absence) of HTN appeared to decrease (increase) the likelihood of mortality.

Across all models, the 5 comorbidities most consistently with the greatest influence on mortality prediction were: CKD, advanced age (>80 years), elevated PTH (>6.8pmol/L), cardiovascular disease (CAD, MI, AF or HTN) and PRCF residence.

3. Discussion

Seven ML models were derived to predict in-hospital mortality for hospitalized elderly minimal trauma HF patients using only categorical data and their performances compared. Overall, the models had reasonable to good performance. An analysis of each model and application SHAP analysis was also performed to gain insight into feature importance.

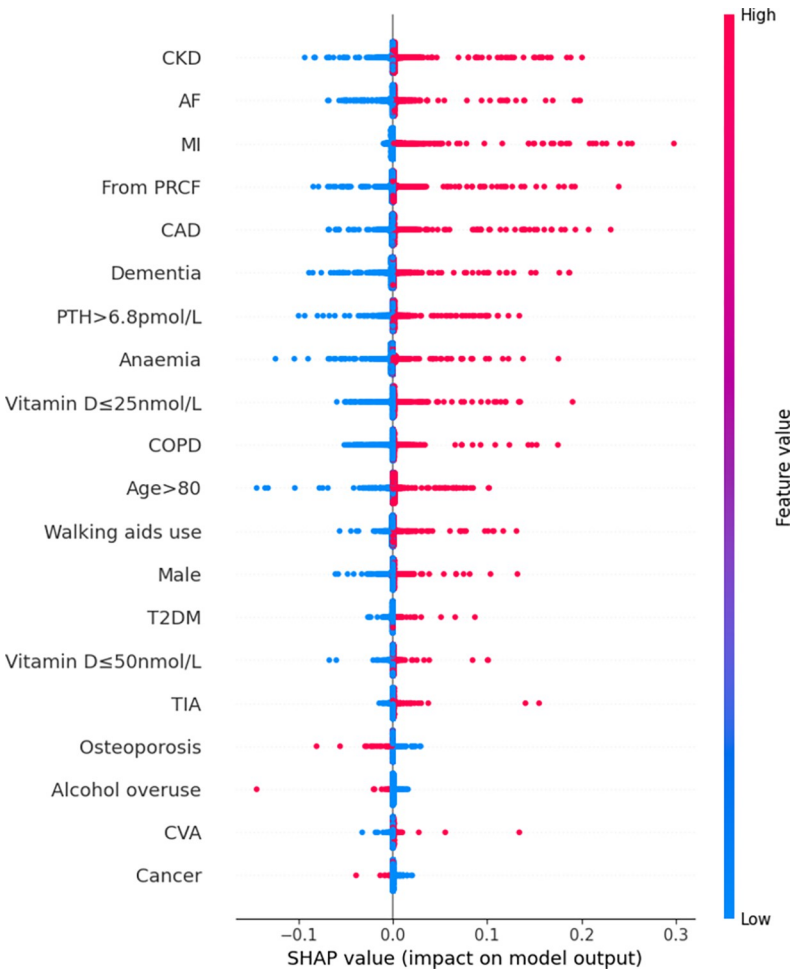

Fig 3. Summary plot SHAP values for patient comorbidities–Naïve Bayes.

https://doi.org/10.1371/journal.pdig.0000529.g003

3.1. Model performance–training and test

Notably, but unsurprisingly, classification performance showed some variation among algorithms. The trained models, ordered in decreasing performance (based on both test AUROCs and test AUPRCs), were RF, NB, XGB and LR (all with no statistically significant difference in performance–see Table 4, Table 5 and S1 Fig) followed by MLP and DT (no statistically significant difference in performance) and finally SVM. AUROCs ranged from 0.500 (SVM) to almost 0.700 (good performance), while AUPRC values ranged from 0.050 (SVM) to 0.115; a reflection of using a simplified model (with binary input data) to perform predictions on a minority class in this imbalanced dataset. There was minimal difference between the training and cross-validation performance for the top 4 models (RF, NB, XGB and LR). A greater variation in training and cross-validation performance scores was noted for DT and MLP, an indicator of overtraining (an infamous tendency in machine learning). That overtraining has occurred despite systematic and meticulous hyperparameter tuning, is strongly suggestive of insufficient data.

To the investigator's knowledge, most studies have focused on only training and applying one class of machine learning algorithm. Often there is no baseline model trained using traditional statistics (e.g. LR). Indeed, most studies have solely utilized tree-based methods (e.g. applying DT, XGB and RF methods) and this is reflected in a scoping study of ML usage in

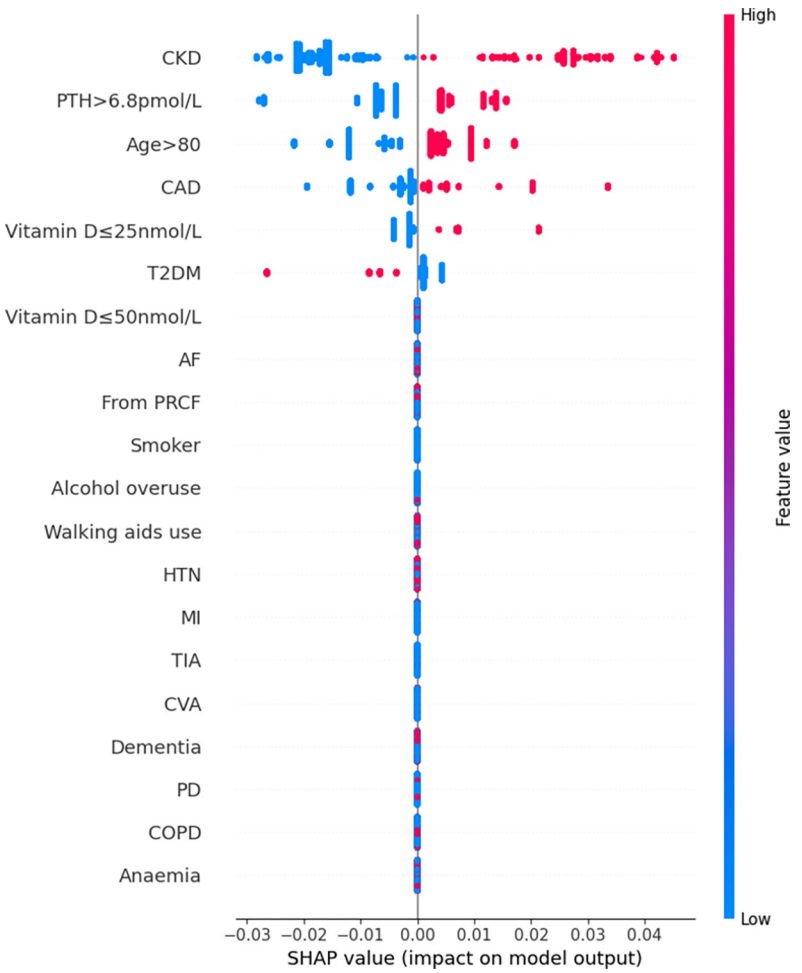

Fig 4. Summary plot SHAP values for patient comorbidities–Decision Tree.

https://doi.org/10.1371/journal.pdig.0000529.g004

health economics and research (on 805 studies) which found the most frequent algorithms used were tree-based methods followed by regression-based (linear/logistic) methods, SVM, NN and finally NB [35]. However, it is known that performance on various tasks varies with different ML algorithms[36] and the finding that predictive performance varies among machine learning algorithms (for the same problem, using the same data) is consistent with this. It is thus ideal that in future applications of machine learning, a more comprehensive set of algorithms are trained, or some justification should be provided, if possible, when certain algorithms are not included.

The performance of NB in predicting mortality is on par with RF, XGB and LR which warrants further discussion here as it has received relatively little attention in the literature. Key to its success is the simplifying assumption of conditional independence among all patient input features. The most obvious advantage from this is that, by virtue of such a simplification, it is computationally inexpensive and is fast to train and run. However, with such a large, seemingly excessive, assumption (that is not strictly satisfied in the current database, with interdependent features such as vitamin D insufficiency and deficiency), it may seem surprising that this model performs so well. Contrary to intuition, its good performance is not a coincidental or even unexpected phenomenon; formal analysis of NBs has established it performs well

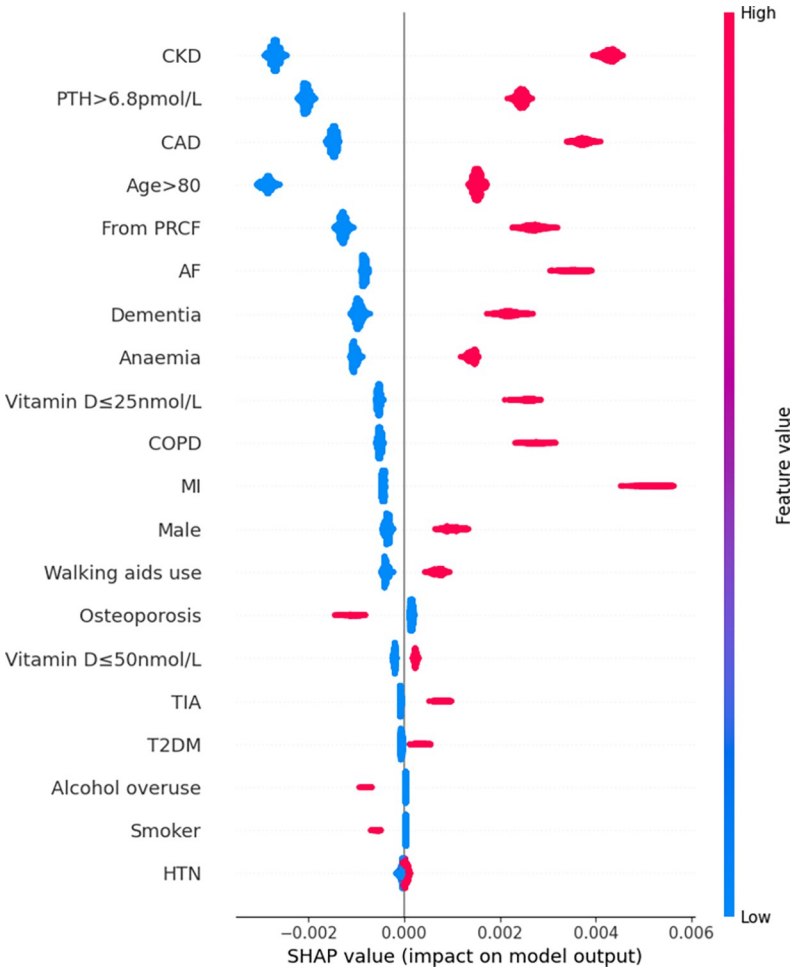

Fig 5. Summary plot SHAP values for patient comorbidities–Random Forest.

https://doi.org/10.1371/journal.pdig.0000529.g005

because the interdependencies, when they do exist, occur in a manner which results in them 'cancel[ling] each other out' and impact only probability estimates, not overall classification performance [37].

3.2. Feature importance–model interpretation and SHAP analysis

Rankings of patient comorbidity importance in their role in mortality prediction were determined from all models from direct interpretation of feature coefficients (see Fig 1 and Table C in S3 Appendix). CKD was most consistently ranked as one of the 5 most important patient comorbidities in predicting mortality. The other most important patient features included markers reflective of bone metabolism (PTH, vitamin D levels) and cardiovascular disease (presence of either one of CAD, MI, AF). Similar trends were found via SHAP value analyses for each model, i.e. CKD, bone metabolism markers and presence of cardiovascular diseases had the strongest influence on prediction of mortality based on mean SHAP values (Figs 2–7). It is recognized in the literature that cardiovascular comorbidities and renal function are important for prognostication which is reflected in their inclusion as input parameters for non-cardiac surgery risk assessment tools such as the Revised Cardiac Risk Index and the American College of Surgeons—surgical risk calculator [38–42]. However, these features are

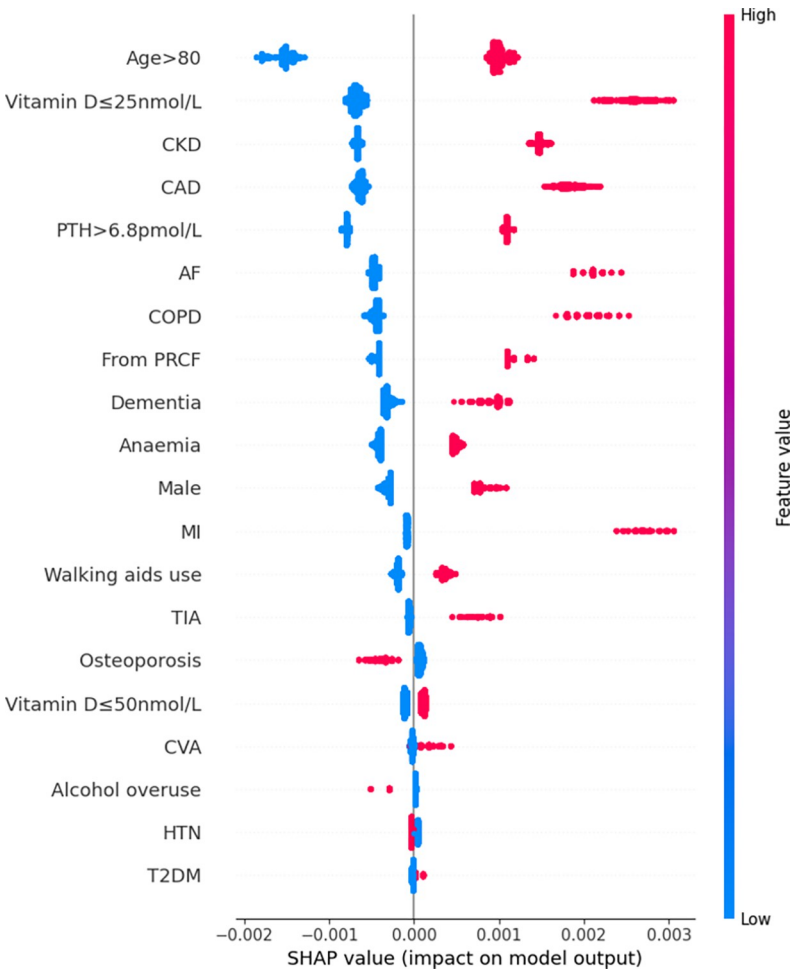

Fig 6. Summary plot SHAP values for patient comorbidities–Extreme Gradient Boosting.

https://doi.org/10.1371/journal.pdig.0000529.g006

not explicitly included in HF-specific risk assessment tools (e.g. in O-POSSUM only symptoms and clinical findings suggestive of cardiovascular disease are included and NHFS only the number of comorbidities is included as an input parameter) [7–10]. Moreover, neither PTH or vitamin D levels are included in any of the current tools, despite an increasing number of studies supporting the key role they play in bone metabolism and prevention of fracture [43–55] and, potentially, with increasing recognition of their importance in the immunity [56–58] prevention of post-operative complications such as hospital acquired infections.

3.3. Further insights from model analysis

Of the four most predictive models, NB and LR models offer intuitive, quantifiable insights into feature contributions to prediction: in LR, the odds ratio can be taken by calculating the exponent of the coefficients, while in NB, from the method of scoring input features (see S1 Appendix), each coefficient corresponds to the ratio of the rate of the comorbidity in those who experienced in-hospital mortality compared to the comorbidity rate in those who survived. So, for example, in predicting mortality, one can see from the LR model that CAD, with a score of 0.319 (95%CI 0.180–0.458) increased mortality risk by 37% (OR 1.37; 95%CI 1.20–1.58) and CKD, with a score of 0.711 (95%CI 0.505–0.918) increased mortality risk by 2.03

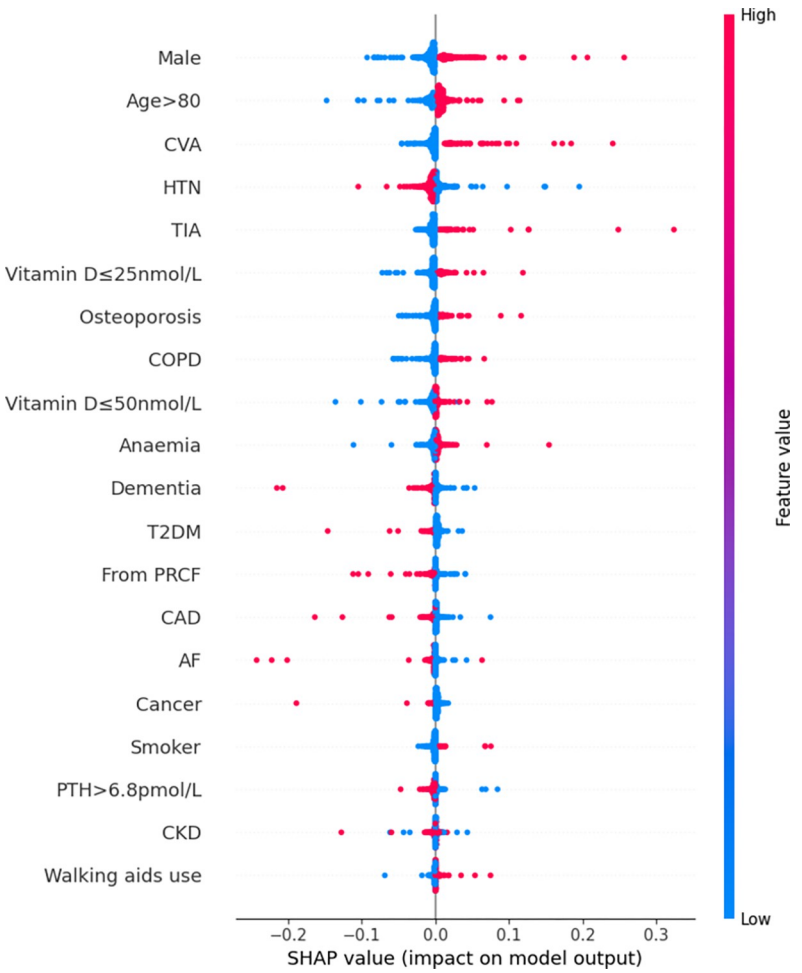

Fig 7. Summary plot SHAP values for patient comorbidities–Multi-Layer Perceptron.

https://doi.org/10.1371/journal.pdig.0000529.g007

(95%CI 1.66–2.50). From the NB model, a score of 1.62 (for CAD), and a score of 1.67 (for CKD) indicated that the rate of each comorbidity was greater in mortality than in survival by 62% and 67% respectively.

For the other top predictive models, insights gained from direct interpretation of RF and XGB is not so straightforward. Both these methods are based on DTs, which is itself an interpretable and intuitive model. However, a major drawback of DTs is that they are very prone to bias and variance (overfitting). RF and XGB address this issue by constructing multiple DTs and the overall prediction is then made from an ensemble/collection of multiple trees (numbering in the hundreds) and, hence, increased predictive performance is obtained at the expense of interpretability. In this study, the coefficients for each feature correspond to the relatively abstract concept of mean decrease in (Gini) impurity (see S1 Appendix).

3.4. Further insights from SHAP values

SHAP values revealed that the presence of more 'severe' comorbidities in each ML model had a more important additive effect on mortality risk than less severe comorbidities, as one might expect. For instance, patients with a history of acute MI (a higher severity sub cohort of CAD patients) typically had the greatest SHAP value indicating that the presence of history of past

MI had the greatest additive effect on mortality prediction. Similarly, the presence of vitamin D deficiency (≤25nmol/L) was correlated with greater SHAP values compared to vitamin D insufficiency (≤50nmol/L) (see Figs 2–7). In contrast, the absence of both MI and vitamin D deficiency in patients had less of a negative effect on mortality prediction compared to the other comorbidities (hence explaining their lower overall importance based on mean SHAP values). This likely is a reflection of their relatively low prevalence in the cohort (Table 1).

For LR, RF, and XGB the SHAP values had low variability and were highly concentrated–an indication that the corresponding input patient features were consistently strong contributors to mortality prediction; a corollary of this was that these models offered good population level insight into mortality risk. Of the top four models, NB was the only model in which the SHAP values themselves varied among individuals. This variability in SHAP values among patients suggested that the influence of each singular comorbidity was not constant, and that each prediction appeared to be tailored toward individual.

3.5. Clinical implications

This NB model could direct individualised HF prevention measures for patients in the community. Currently there exist osteoporotic fracture risk assessment tools, such as the FRAX and the Garvan tools [59–62], however, a reliable method to identify those at a high risk of in-hospital mortality remains elusive. A quantifiable and objective prognostic estimate of mortality risk following HF would guide clinicians in, firstly, triaging referrals to falls prevention clinics and, secondly, objectively appraise the need to commence anti-osteoporotic agents (based on in-hospital mortality risk following HF) against the risk of drug-related adverse effects [63,64]. The tool can also aid in identifying patients at risk of mortality early during their admission. It has been established that early surgical intervention reduces mortality risk in elderly HF patient admissions [65–67]. The model could be used by clinicians to assist in prioritisation of surgical intervention for those HF patients identified as high-risk for in-hospital mortality. By excluding the need for laboratory parameters, the ML model can be used by emergency physicians, orthopaedic surgeons to prioritise surgery for patients classified as high-risk with minimal delay. It can also prompt clinicians to set the expectations of patients and next of kin in the early phases of admission. Early discussions and clear communication with patients and family are a key element of clinical care and would facilitate better preparedness for end-of-life care in the event of rapid in-hospital deterioration, minimizing miscommunication, dissatisfaction and bereavement while maximising quality of life [68–70].

That the developed model does not depend on laboratory or intra-operative data confers a significant advantage over both the NHFS and O-POSSUM–it can be applied much earlier than either. The variables required for the calculation of NHFS are: age (<66, 66–85, ≥86 years), sex, admission haemoglobin, mini-mental test score, living in an institution, number of comorbidities (≥ 2), and presence of active malignancy [7,8]. The variables required for the calculation of the O-POSSUM are: age (in years), chest radiograph findings, respiratory symptoms, cardiac signs, vital signs (systolic blood pressure in mmHg, pulse in beats/min), Glasgow Coma Scale, full blood count (haemoglobin, white cell count), electrolytes (sodium, potassium, urea), electrocardiogram findings, operative severity, multiple procedures, total blood loss, peritoneal soiling, presence of malignancy, mode of surgery (emergent vs elective) [10]. For the NHFS, haemoglobin on admission is not immediately available, and the mini-mental test assessment, while possible to perform at the bedside, is not routinely performed as part of the initial assessment for a HF admission. For the O-POSSUM score, not only does it require laboratory test results (electrolytes, full blood count), but the status of certain features (i.e. respiratory symptoms, cardiac signs, chest radiograph, electrocardiogram) are victim to subjective

(clinician dependent) interpretation and, finally, intraoperative data is required for the O-POSSUM model which precludes its potential use in the outpatient setting and early phases of hospital admission. While the NB model can clearly be applied in a more timely manner than either NHFS or O-POSSUM, a comparison of their discriminatory performances was not possible in this study as, due to limitations on the dataset, it was not possible to calculate the corresponding patient scores.

The results from the analysis of feature importance (both SHAP and direct approach) can also be used to guide treatment in the context of osteoporotic HF. Certain comorbidities, such as CKD, have been found in this study to have large contributions to patient mortality prediction across trained ML models. These findings could advance current clinical practice, if validated externally; they support early implementation of HF prevention strategies as standard of care for patients with these select, highly influential, comorbidities. On a case-by-case basis, SHAP values also offer local explanations, i.e. explanations specific to individualised patients and predictions, enabling stakeholders (primarily patients and their clinicians) to make informed decisions, expose underlying vulnerabilities and protect individuals from the potential pitfalls of automated decisions.

3.6. Limitations and future work

Internal nested cross-validation, though relatively rigorous, is no substitute for external validation. The initial step toward this would be temporal validation; patients meeting the same criteria as those defined in the 'Methods–Data Collection' section will have data collected in the period following 2019 and this data will be used to validate the developed algorithm. Following temporal validation, the aim would be to validate on a wider geographic region (inter-hospital, to inter-state and potentially international cohorts). Following rigorous validation, there remains the challenge of model dissemination and integration into clinical practice. A possible approach would be to implement the model into a web-based application making it readily accessible to any healthcare provider with internet access, though such an approach could result in improper use of the algorithm on non-validated populations. Alternatively, the model could be integrated into commercial Electronic Health Records used by hospitals, though this would be at the expense of limiting users to those with access to specific (proprietary) software.

It should also be noted that this was a study on a retrospective cohort, with members recruited from a single-centre. Though the dataset used here is not unreasonably small, it must be acknowledged that it may still be insufficient: firstly, because of the overfitting noted in MLP models and secondly because of the imbalance inherent to the dataset with a 5% mortality rate. With only 189 cases, the mortality population may be under-represented from a machine-learning perspective (which typically requires cohort sizes numbering in the 1000s or greater to be trained effectively). Moreover, inaccurate reports (probable under-reporting) on smoking and drinking habits by patients may bias findings. Furthermore, analysis and model derivation has been conducted using only categorical features which may negatively impact predictive ability. Model predictions were not calibrated, and it is known that certain machine learning models, particularly NB are notoriously poor at estimating probabilities despite being good classifiers.

3.7. Conclusion and final comments

In summary, NB was the most optimal model having the optimal virtues of strong predictive performance, model interpretability and potential for making individualized predictions. While RF, XGB and LR had similar performance capabilities, by nature they are not readily interpretable (i.e. RF and XGB) or are not optimal for individualized predictions (i.e. LR).

With ongoing development of digital infrastructure in the healthcare industry it is inevitable that machine learning algorithms will only become increasingly powerful and commonplace. As we await this reality, it is to be hoped that the findings here will provide physicians and clinicians with a tool that can be used to rapidly identify patients at higher risk of mortality early by knowledge of patient comorbidities; currently most prognostication tools can only be applied later in the admission. Moreover, hopefully this study provides valuable insights in applying ML models in healthcare for clinicians and researchers, in particular the advantages of the computationally inexpensive NB models highlighting its simplicity and interpretability with negligible compromise in performance.

4. Materials and methods

4.1. Ethics statement

The study was conducted in accordance with the Declaration of Helsinki (1964) and the Council for International Organisations of Medical Sciences International Ethic Guidelines and approved by the Australian Capital Territory Human Research Ethics Committee on the 31st May 2023 (reference number: 2023.LRE.00063). Patients' written informed consent was waived because analysis was performed on a digital anonymised database.

4.2. Data collection

Our cohort comprised 3625 elderly (i.e. aged $\geq$ 65 years of age) patients consecutively admitted to the Department of Orthopaedic Surgery at the Canberra Hospital between 1999–2019 with osteoporotic hip fracture. Patients admitted with hip fracture secondary to moderate-high energy trauma, or secondary to minimal trauma but with malignancy associated pathological fracture were excluded. Data on in-hospital mortality, sociodemographic features (age, sex, smoking status, active history of overuse of alcohol, use of walking aids, and if the patient was a resident of an permanent residential care facility [PRCF]) and comorbidities (presence of hypertension [HTN], coronary artery disease [CAD], previous history of acute myocardial infarction [MI], atrial fibrillation [AF], past history of stroke [cerebrovascular accident, CVA], transient ischaemic attack [TIA], dementia, Parkinson's disease [PD], chronic obstructive pulmonary disease [COPD], type 2 diabetes mellitus [T2DM], chronic kidney disease [CKD], anaemia, history of solid organ malignancy, osteoporosis and hyperparathyroidism [parathyroid hormone/PTH>6.8pmol/L] and vitamin D insufficiency/deficiency; (25)OH vitamin D $\leq$ 50/25nmol/L) were collected. Binary variables were assumed to follow a Bernoulli distribution. For parathyroid hormone, the upper limit of the laboratory reference range (6.8 pmol/L) was chosen as the cutoff. The definition of vitamin D insufficiency and deficiency was based on those utilised previously in the literature [71,72].

The continuous variable of age was tested for normality (via visualisation using histogram, the Kolmogorov-Smirnov test, Shapiro-Wilke test and the Anderson-Darling test). Results suggested a non-normal distribution–hence median and interquartile ranges were used as measures of central tendency and spread respectively. The age cutoff was obtained by taking the median age of the cohort (84 years) and rounding to the nearest decade to 80 years.

There were few instances of missing data (see Table B in S3 Appendix). For this reason, patients with missing data were omitted from analysis.

4.3. Model development

Seven ML algorithms (LR, SVM, NB, DT, RF, XGB and MLP) were trained to predict mortality. For each of the algorithms, model selection and evaluation were performed using nested

cross-validation. For each iteration of k-fold cross validation, data was shuffled and split into k stratified subsets (i.e. class proportions for mortality were maintained across all data partitions); this was performed due to the significant imbalance in the dataset. The same random seed was used for data shuffling in the development of each ML model.

Model selection was performed on the inner, 3-fold, cross-validation loop; a grid search on key hyperparameters (identified using an approach based on the fractional factorial design of experiments) was used to identify the optimal hyperparameter configuration using the mean area under the receiver operating characteristic (AUROC) on the validation set as the performance criteria. Computations were performed using the Python packages, sklearn and pandas [73,74].

4.4. Model performance (and comparisons)

Model performance was evaluated on the outer (5-fold) cross-validation loop using the mean of the validation AUROCs. The student t-test was used to compare mean AUROCs of different ML models against one other. Additionally, given the imbalance to the dataset, the mean area under the precision-recall curve (AUPRC) for performance on the validation set was calculated for each trained ML. Computations were performed using the Python package SciPy (in particular 'scipy.stats' routines) [75].

4.5. Feature importance–model interpretation

Each trained model was analysed directly. In general, the training of each model involved optimization of coefficients corresponding to each patient feature (comorbidity). The trained models were analysed; for each patient comorbidity a corresponding coefficient or score was computed (see S1 Appendix). Features were ranked by importance based on the values of these scores.

4.6. Feature importance–SHAP analysis

For each patient, the SHAP value allocates a quantifiable credit to each variable (i.e. patient comorbidity) in its contribution to the model output (i.e. the final prediction). Feature importance analysis with SHAP was performed using the Python implementation–further details can be found in S2 Appendix [30,33]. Features were ranked based on the mean SHAP values for each comorbidity.

Supporting information

S1 Fig. ML model test performance (area under the receiver operating characteristic, AUROC). Only the test set AUCs evaluated from the 5-fold cross-validation for the four best-performing ML models are shown.
(PPTX)

S1 Appendix. Coefficient analysis.
(DOCX)

S2 Appendix. SHAP analysis.
(DOCX)

S3 Appendix. Supplementary Data.
(DOCX)

Author Contributions

Conceptualization: Jo-Wai Douglas Wang.

Data curation: Jo-Wai Douglas Wang.

Formal analysis: Jo-Wai Douglas Wang.

Investigation: Jo-Wai Douglas Wang.

Methodology: Jo-Wai Douglas Wang.

Software: Jo-Wai Douglas Wang.

Validation: Jo-Wai Douglas Wang.

Visualization: Jo-Wai Douglas Wang.

Writing – original draft: Jo-Wai Douglas Wang.

Writing – review & editing: Jo-Wai Douglas Wang.

References

1. Veronese N, Maggi S. Epidemiology and social costs of hip fracture. Injury. 2018; 49(8):1458–60. https://doi.org/10.1016/j.injury.2018.04.015 PMID: 29699731
2. White SM, Griffiths R. Projected incidence of proximal femoral fracture in England: a report from the NHS Hip Fracture Anaesthesia Network (HIPFAN). Injury. 2011; 42(11):1230–3. https://doi.org/10.1016/j.injury.2010.11.010 PMID: 21183180
3. Groff H, Kheir MM, George J, Azboy I, Higuera CA, Parvizi J. Causes of in-hospital mortality after hip fractures in the elderly. Hip Int. 2020; 30(2):204–9. https://doi.org/10.1177/1120700019835160 PMID: 30909746
4. Sheehan KJ, Sobolev B, Guy P, Kuramoto L, Morin SN, Sutherland JM, et al. In-hospital mortality after hip fracture by treatment setting. Cmaj. 2016; 188(17–18):1219–25. https://doi.org/10.1503/cmaj.160522 PMID: 27754892
5. Hip fracture incidence and hospitalisations in Australia 2015–16. Canberra: AIHW: Australian Institute of Health and Welfare 2018.
6. Wu TY, Jen MH, Bottle A, Liaw CK, Aylin P, Majeed A. Admission rates and in-hospital mortality for hip fractures in England 1998 to 2009: time trends study. J Public Health (Oxf). 2011; 33(2):284–91. https://doi.org/10.1093/pubmed/fdq074 PMID: 20926392
7. Sun L, Liu Z, Wu H, Liu B, Zhao B. Validation of the Nottingham Hip Fracture Score in Predicting Postoperative Outcomes Following Hip Fracture Surgery. Orthop Surg. 2023; 15(4):1096–103. https://doi.org/10.1111/os.13624 PMID: 36794402
8. Olsen F, Lundborg F, Kristiansson J, Hård Af Segerstad M, Ricksten SE, Nellgård B. Validation of the Nottingham Hip Fracture Score (NHFS) for the prediction of 30-day mortality in a Swedish cohort of hip fractures. Acta Anaesthesiol Scand. 2021; 65(10):1413–20. https://doi.org/10.1111/aas.13966 PMID: 34363201
9. Jones HJ, de Cossart L. Risk scoring in surgical patients. Br J Surg. 1999; 86(2):149–57. https://doi.org/10.1046/j.1365-2168.1999.01006.x PMID: 10100780
10. Mohamed K, Copeland GP, Boot DA, Casserley HC, Shackleford IM, Sherry PG, et al. An assessment of the POSSUM system in orthopaedic surgery. J Bone Joint Surg Br. 2002; 84(5):735–9. https://doi.org/10.1302/0301-620x.84b5.12626 PMID: 12188495
11. Hill BL, Brown R, Gabel E, Rakocz N, Lee C, Cannesson M, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth. 2019; 123(6):877–86. https://doi.org/10.1016/j.bja.2019.07.030 PMID: 31627890
12. Hu XY, Liu H, Zhao X, Sun X, Zhou J, Gao X, et al. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. CNS Neurosci Ther. 2022; 28(4):608–18. https://doi.org/10.1111/cns.13758 PMID: 34792857
13. Bishara A, Chiu C, Whitlock EL, Douglas VC, Lee S, Butte AJ, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. BMC Anesthesiol. 2022; 22(1):8. https://doi.org/10.1186/s12871-021-01543-y PMID: 34979919

14. Zhang J, Jiang L, Zhu X. A Machine Learning-Modified Novel Nomogram to Predict Perioperative Blood Transfusion of Total Gastrectomy for Gastric Cancer. Front Oncol. 2022; 12:826760. https://doi.org/10.3389/fonc.2022.826760 PMID: 35480095

15. Peng X, Zhu T, Wang T, Wang F, Li K, Hao X. Machine learning prediction of postoperative major adverse cardiovascular events in geriatric patients: a prospective cohort study. BMC Anesthesiol. 2022; 22(1):284. https://doi.org/10.1186/s12871-022-01827-x PMID: 36088288

16. Neto PCS, Rodrigues AL, Stahlschmidt A, Helal L, Stefani LC. Developing and validating a machine learning ensemble model to predict postoperative delirium in a cohort of high-risk surgical patients: A secondary cohort analysis. Eur J Anaesthesiol. 2023; 40(5):356–64. https://doi.org/10.1097/EJA.0000000000001811 PMID: 36860180

17. Forssten MP, Bass GA, Ismail AM, Mohseni S, Cao Y. Predicting 1-Year Mortality after Hip Fracture Surgery: An Evaluation of Multiple Machine Learning Approaches. J Pers Med. 2021; 11(8). https://doi.org/10.3390/jpm11080727 PMID: 34442370

18. Li YY, Wang JJ, Huang SH, Kuo CL, Chen JY, Liu CF, et al. Implementation of a machine learning application in preoperative risk assessment for hip repair surgery. BMC Anesthesiol. 2022; 22(1):116. https://doi.org/10.1186/s12871-022-01648-y PMID: 35459103

19. Lei M, Han Z, Wang S, Han T, Fang S, Lin F, et al. A machine learning-based prediction model for in-hospital mortality among critically ill patients with hip fracture: An internal and external validated study. Injury. 2023; 54(2):636–44. https://doi.org/10.1016/j.injury.2022.11.031 PMID: 36414503

20. Zhao H, You J, Peng Y, Feng Y. Machine Learning Algorithm Using Electronic Chart-Derived Data to Predict Delirium After Elderly Hip Fracture Surgeries: A Retrospective Case-Control Study. Front Surg. 2021; 8:634629. https://doi.org/10.3389/fsurg.2021.634629 PMID: 34327210

21. Metsis V, Androutsopoulos I, Paliouras G. Spam Filtering with Naive Bayes—Which Naive Bayes? Conference on Email and Anti-Spam; Mountain View, California USA2006.

22. Blanco R, Inza I, Merino M, Quiroga J, Larrañaga P. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. Journal of Biomedical Informatics. 2005; 38 (5):376–88. https://doi.org/10.1016/j.jbi.2005.05.004 PMID: 15967731

23. Galiatsatos D, Anastassopoulos G, Drosos G, Ververidis A, Tilkeridis K, Kazakos K. Prediction of 30-Day Mortality after a Hip Fracture Surgery Using Neural and Bayesian Networks2014. 566–75 p.

24. Cui S, Zhao L, Wang Y, Dong Q, Ma J, Wang Y, et al. Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. Injury. 2018; 49(10):1865–70. https://doi.org/10.1016/j.injury.2018.07.025 PMID: 30097310

25. Yun K, Oh J, Hong TH, Kim EY. Prediction of Mortality in Surgical Intensive Care Unit Patients Using Machine Learning Algorithms. Front Med (Lausanne). 2021; 8:621861. https://doi.org/10.3389/fmed.2021.621861 PMID: 33869245

26. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Computing and Applications. 2020; 32(24):18069–83. https://doi.org/10.1007/s00521-019-04051-w

27. Lu SC, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Front Oncol. 2023; 13:1129380. https://doi.org/10.3389/fonc.2023.1129380 PMID: 36925929

28. Sathyan A, Weinberg AI, Cohen K. Interpretable AI for bio-medical applications. Complex Eng Syst. 2022; 2(4). https://doi.org/10.20517/ces.2022.41 PMID: 37025127

29. Ejiyi CJ, Qin Z, Amos J, Ejiyi MB, Nnani A, Ejiyi TU, et al. A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms. Healthcare Analytics. 2023; 3:100166. https://doi.org/10.1016/j.health.2023.100166

30. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TWV, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Scientific Reports. 2021; 11(1):23017. https://doi.org/10.1038/s41598-021-02481-y PMID: 34837021

31. Tang S, Ghorbani A, Yamashita R, Rehman S, Dunnmon JA, Zou J, et al. Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. Scientific Reports. 2021; 11(1):8366. https://doi.org/10.1038/s41598-021-87762-2 PMID: 33863957

32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020; 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9 PMID: 32607472

33. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Neural Information Processing Systems; Long Beach, CA, USA2017.

34. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems. 2014; 41(3):647–65. https://doi.org/10.1007/s10115-013-0679-x

35. Lee W, Schwartz N, Bansal A, Khor S, Hammarlund N, Basu A, et al. A Scoping Review of the Use of Machine Learning in Health Economics and Outcomes Research: Part 2-Data From Nonwearables. Value Health. 2022; 25(12):2053–61. https://doi.org/10.1016/j.jval.2022.07.011 PMID: 35989154

36. Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd international conference on Machine learning—ICML '06. 2006;2006:161–8. https://doi.org/10.1145/1143844.1143865

37. Zhang H. The Optimality of Naive Bayes. Florida Artificial Intelligencce Research Society; Palo Alto, California USA2004.

38. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013; 217(5):833–42.e1-3. https://doi.org/10.1016/j.jamcollsurg.2013.07.385 PMID: 24055383

39. Edelstein AI, Kwasny MJ, Suleiman LI, Khakhkhar RH, Moore MA, Beal MD, et al. Can the American College of Surgeons Risk Calculator Predict 30-Day Complications After Knee and Hip Arthroplasty? J Arthroplasty. 2015; 30(9 Suppl):5–10. https://doi.org/10.1016/j.arth.2015.01.057 PMID: 26165953

40. Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. Ann Intern Med. 2010; 152(1):26–35. https://doi.org/10.7326/0003-4819-152-1-201001050-00007 PMID: 20048269

41. Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B, et al. Multifactorial index of cardiac risk in noncardiac surgical procedures. N Engl J Med. 1977; 297(16):845–50. https://doi.org/10.1056/NEJM197710202971601 PMID: 904659

42. Goltz DE, Baumgartner BT, Politzer CS, DiLallo M, Bolognesi MP, Seyler TM. The American College of Surgeons National Surgical Quality Improvement Program Surgical Risk Calculator Has a Role in Predicting Discharge to Post-Acute Care in Total Joint Arthroplasty. J Arthroplasty. 2018; 33(1):25–9. https://doi.org/10.1016/j.arth.2017.08.008 PMID: 28899592

43. Ban Z-N, Li Z-J, Gu Q-S, Cheng J, Huang Q, Xing S-X. Correlation of serum PTH level and fracture healing speed in elderly patients with hip fracture. Journal of Orthopaedic Surgery and Research. 2019; 14(1):361. https://doi.org/10.1186/s13018-019-1413-5 PMID: 31718681

44. Byun SE, Lee S, Kim JW, Ha YC, Kim CH, Ha C, et al. Preventive Effects of Low Parathyroid Hormone Levels on Hip Fracture in Patients with Vitamin D Deficiency. J Bone Metab. 2019; 26(2):89–95. https://doi.org/10.11005/jbm.2019.26.2.89 PMID: 31223605

45. Caffarelli C, Mondanelli N, Crainz E, Giannotti S, Frediani B, Gonnelli S. The Phenotype of Bone Turnover in Patients with Fragility Hip Fracture: Experience in a Fracture Liaison Service Population. Int J Environ Res Public Health. 2022; 19(12). https://doi.org/10.3390/ijerph19127362 PMID: 35742610

46. Han J, Cho Y, Jee S, Jo S. Vitamin D Levels in Patients with Low-energy Hip Fractures. Hip Pelvis. 2020; 32(4):192–8. https://doi.org/10.5371/hp.2020.32.4.192

47. Kanis JA, Harvey NC, Liu E, Vandenput L, Lorentzon M, McCloskey EV, et al. Primary hyperparathyroidism and fracture probability. Osteoporos Int. 2023; 34(3):489–99. https://doi.org/10.1007/s00198-022-06629-y PMID: 36525071

48. LeBoff MS, Chou SH, Ratliff KA, Cook NR, Khurana B, Kim E, et al. Supplemental Vitamin D and Incident Fractures in Midlife and Older Adults. New England Journal of Medicine. 2022; 387(4):299–309. https://doi.org/10.1056/NEJMoa2202106 PMID: 35939577

49. Ng K, St. John A, Bruce DG. Secondary hyperparathyroidism, vitamin D deficiency and hip fracture: importance of sampling times after fracture. Bone and Mineral. 1994; 25(2):103–9. https://doi.org/10.1016/S0169-6009(08)80252-8 PMID: 8086849

50. Seib CD, Meng T, Suh I, Harris AHS, Covinsky KE, Shoback DM, et al. Risk of Fracture Among Older Adults With Primary Hyperparathyroidism Receiving Parathyroidectomy vs Nonoperative Management. JAMA Internal Medicine. 2022; 182(1):10–8. https://doi.org/10.1001/jamainternmed.2021.6437 PMID: 34842909

51. Tan LYJ, Asri NAM. ODP097 Hyperparathyroidism as a major risk factor of fracture and its prevalence in patients with hip fracture. Journal of the Endocrine Society. 2022;6(Supplement_1):A165–A6. https://doi.org/10.1210/jendso/bvac150.340

52. Wang N, Chen Y, Ji J, Chang J, Yu S, Yu B. The relationship between serum vitamin D and fracture risk in the elderly: a meta-analysis. Journal of Orthopaedic Surgery and Research. 2020; 15(1):81. https://doi.org/10.1186/s13018-020-01603-y PMID: 32103764

53. Wang Q, Yu D, Wang J, Lin S. Association between vitamin D deficiency and fragility fractures in Chinese elderly patients: a cross-sectional study. Annals of Palliative Medicine. 2020; 9(4):1660–5. https://doi.org/10.21037/apm-19-610 PMID: 32527135

54. Yakabe M, Hosoi T, Matsumoto S, Fujimori K, Tamaki J, Nakatoh S, et al. Prescription of vitamin D was associated with a lower incidence of hip fractures. Scientific Reports. 2023; 13(1):12889. https://doi.org/10.1038/s41598-023-40259-6 PMID: 37558795

55. Yao P, Bennett D, Mafham M, Lin X, Chen Z, Armitage J, et al. Vitamin D and Calcium for the Prevention of Fracture: A Systematic Review and Meta-analysis. JAMA Network Open. 2019; 2(12):e1917789-e. https://doi.org/10.1001/jamanetworkopen.2019.17789 PMID: 31860103

56. Fernandez GJ, Ramírez-Mejía JM, Urcuqui-Inchima S. Vitamin D boosts immune response of macrophages through a regulatory network of microRNAs and mRNAs. The Journal of Nutritional Biochemistry. 2022; 109:109105. https://doi.org/10.1016/j.jnutbio.2022.109105 PMID: 35858666

57. Sinder BP, Pettit AR, McCauley LK. Macrophages: Their Emerging Roles in Bone. J Bone Miner Res. 2015; 30(12):2140–9. https://doi.org/10.1002/jbmr.2735 PMID: 26531055

58. Small AG, Harvey S, Kaur J, Putty T, Quach A, Munawara U, et al. Vitamin D upregulates the macrophage complement receptor immunoglobulin in innate immunity to microbial pathogens. Communications Biology. 2021; 4(1):401. https://doi.org/10.1038/s42003-021-01943-3 PMID: 33767430

59. Trémollieres FA, Pouillès JM, Drewniak N, Laparra J, Ribot CA, Dargent-Molina P. Fracture risk prediction using BMD and clinical risk factors in early postmenopausal women: sensitivity of the WHO FRAX tool. J Bone Miner Res. 2010; 25(5):1002–9. https://doi.org/10.1002/jbmr.12 PMID: 20200927

60. Agarwal A, Leslie WD, Nguyen TV, Morin SN, Lix LM, Eisman JA. Performance of the Garvan Fracture Risk Calculator in Individuals with Diabetes: A Registry-Based Cohort Study. Calcif Tissue Int. 2022; 110(6):658–65. https://doi.org/10.1007/s00223-021-00941-1 PMID: 34994831

61. Watts NB. The Fracture Risk Assessment Tool (FRAX): applications in clinical practice. J Womens Health (Larchmt). 2011; 20(4):525–31. https://doi.org/10.1089/jwh.2010.2294 PMID: 21438699

62. Agarwal A, Leslie WD, Nguyen TV, Morin SN, Lix LM, Eisman JA. Predictive performance of the Garvan Fracture Risk Calculator: a registry-based cohort study. Osteoporos Int. 2022; 33(3):541–8. https://doi.org/10.1007/s00198-021-06252-3 PMID: 34839377

63. Kennel KA, Drake MT. Adverse effects of bisphosphonates: implications for osteoporosis management. Mayo Clin Proc. 2009; 84(7):632–7; quiz 8. https://doi.org/10.1016/S0025-6196(11)60752-0 PMID: 19567717

64. Tay WL, Tay D. Discontinuing Denosumab: Can It Be Done Safely? A Review of the Literature. Endocrinol Metab (Seoul). 2022; 37(2):183–94. https://doi.org/10.3803/EnM.2021.1369 PMID: 35417954

65. Kristiansson J, Hagberg E, Nellgård B. The influence of time-to-surgery on mortality after a hip fracture. Acta Anaesthesiol Scand. 2020; 64(3):347–53. https://doi.org/10.1111/aas.13494 PMID: 31652349

66. Pincus D, Ravi B, Wasserstein D, Huang A, Paterson JM, Nathens AB, et al. Association Between Wait Time and 30-Day Mortality in Adults Undergoing Hip Fracture Surgery. Jama. 2017; 318(20):1994–2003. https://doi.org/10.1001/jama.2017.17606 PMID: 29183076

67. Simunovic N, Devereaux PJ, Sprague S, Guyatt GH, Schemitsch E, Debeer J, et al. Effect of early surgery after hip fracture on mortality and complications: systematic review and meta-analysis. Cmaj. 2010; 182(15):1609–16. https://doi.org/10.1503/cmaj.092220 PMID: 20837683

68. Heyland DK, Barwich D, Pichora D, Dodek P, Lamontagne F, You JJ, et al. Failure to engage hospitalized elderly patients and their families in advance care planning. JAMA Intern Med. 2013; 173(9):778–87. https://doi.org/10.1001/jamainternmed.2013.180 PMID: 23545563

69. Sudore RL, Fried TR. Redefining the "planning" in advance care planning: preparing for end-of-life decision making. Ann Intern Med. 2010; 153(4):256–61. https://doi.org/10.7326/0003-4819-153-4-201008170-00008 PMID: 20713793

70. Wright AA, Zhang B, Ray A, Mack JW, Trice E, Balboni T, et al. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. Jama. 2008; 300(14):1665–73. https://doi.org/10.1001/jama.300.14.1665 PMID: 18840840

71. Holick MF. Vitamin D deficiency. N Engl J Med. 2007; 357(3):266–81. https://doi.org/10.1056/NEJMra070553 PMID: 17634462

72. Fisher A, Goh S, Srikusalanukul W, Davis M. Elevated serum PTH is independently associated with poor outcomes in older patients with hip fracture and vitamin D inadequacy. Calcif Tissue Int. 2009; 85(4):301–9. https://doi.org/10.1007/s00223-009-9283-1 PMID: 19763373

73. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–30.

74. McKinney Wo. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference2010. p. 51–6.

75. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543