Machine Learning for Risk Factor Identification and Cardiovascular Mortality Prediction Among Patients with Osteoporosis

Seyed Alireza Hasheminasab, Daniel Prieto-Alhambra, Marta Pineda Moncusi, Sara Khalid

Abstract— **Risk prediction tools are increasingly popular aids in clinical decision-making. However, the underlying models are often trained on data from general patient cohorts and may not be representative of and suitable for use with targeted patient groups in actual clinical practice, such as in the case of osteoporosis patients who may be at elevated risk of mortality. We developed and internally validated a cardiovascular mortality risk prediction model tailored to individuals with osteoporosis using a range of machine learning models. We compared the performance of machine learning models with existing expert-based models with respect to data-driven risk factor identification, discrimination, and calibration. The proposed models were found to outperform existing cardiovascular mortality risk prediction tools for the osteoporosis population. External validation of the model is recommended.**

Clinical Relevance— **This study presents the performance of machine learning models for cardiovascular death prediction among osteoporotic patients as well as the risk factors identified by the models to be important predictors.**

I. Introduction

Osteoporosis (OP) is a health condition that involves bone deterioration; it can significantly impact a patient's quality of life and can increase the risk of comorbidity and mortality [1]. Due to concerns about the risk of cardiovascular events in patients with osteoporosis, some studies have investigated the risk of cardiovascular disease (CVD) among patients with OP [2, 3]. Findings indicate an increased risk of CVD-related mortality due to myocardial infarction (MI) and stroke in individuals with low bone mineral density, motivating the need for the identification of OP patients at high risk of cardiovascular disease and improving the performance of existing cardiovascular risk prediction tools for OP patients to support clinical decision-making and ultimately enhance patient outcomes [3].

Most CVD risk prediction tools are based on data and assumptions that are compatible with the general population and may not be entirely consistent with OP patient characteristics [4-10]. For instance, although there are common risk factors between CVD and OP such as age, obesity, and type 2 diabetes, some well-known risk factors of osteoporosis such as low weight or female gender can present a protective effect against CVD risk [11-14]. Recently attempts to develop CVD risk models tailored to the OP population have been made using routine clinical data [2].

Authors are with the Centre for Statistics in Medicine, Nuffield Department of Orthopedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, U.K (e-mail: alireza.hasheminasab@ndorms.ox.ac.uk)

The increasing availability of clinical data in the form of large-scale electronic health records has amplified the interest in machine learning (ML) models for clinical risk prediction [15, 16]; with some literature favouring ML over traditional techniques, particularly to account for complex interactions within a wide range of patient characteristics [17-21].

In this paper, we aimed to compare machine-learning models with existing approaches for risk factor identification and risk prediction of CVD death. Proposed models were compared with a reference model based on QRISK - a widely known CVD risk prediction tool used in clinical practice [9, 10]. We further analysed the impact of training sample size on model performance.

II. Methodology

A. Data Source

The Clinical Practice Research Datalink (CPRD) GOLD dataset contains anonymized electronic primary care records for a total of 21 million patients in the UK of which approximately 9 million are eligible for linkage [22]. The dataset includes demographic information, medication prescriptions, clinical history, laboratory tests, and referrals. This study includes a subset of the CPRD dataset including all patients with a diagnosis of osteoporotic from January 1, 1995, to January 31, 2017. To investigate CVD death as the study outcome, we used CPRD data linked with the UK's Office for National Statistics mortality records. We also linked the dataset with the UK National Health Services Hospital Episode Statistics Admitted Patient Care (HES APC) dataset to extract and link in-hospital diagnoses, procedures, and treatments for osteoporosis patients having had hospital episodes. Socio-economic status was included by linkage with the Index of Multiple Deprivation dataset.

B. Study Population

We defined the osteoporosis cohort as the study population, including only patients with Read or ICD-10 codes of an osteoporosis diagnosis [23, 24]. An index date was defined as the date of the first osteoporosis diagnosis. Participants were followed up from the index date up to a maximum of two years. We censored participants at the earliest of the study outcome, non-CVD death, migration/transfer out of GP practice, last date of data availability (based on data extraction date), or end of the two-year follow-up period. We also excluded patients younger than 50 years old and patients with less than one year of data

available prior to the index date. This resulted in a total population of 65,295 osteoporosis patients of whom 2.9% had the study outcome (CVD death).

C. Outcome Definition

CVD death was defined as the occurrence of either stroke or MI recorded as the primary cause of death.

D. Pre-Processing

An initial list of 52 patient characteristics including demographics, socio-economic status, Charlson's comorbidity index, medical history (diagnoses, procedures/operations, laboratory tests) and prescriptions/medications formed the set of candidate features (Figure 1). This set included but was not limited to the 21 QRISK-selected features i.e. CVD risk factors used in the existing QRISK tool.

Missingness within a given characteristic was handled using multiple imputations with chained equations resulting in 20 imputed datasets [25]. Due to computation time constraints, we compared all models on the first resulting imputed set.

1	Sex	2	SES	3	Smoking**
4	Drinking**	5	Diabetes type 1*	6	Diabetes type 2*
7	Chronic obstructive pulmonary disease*	8	Chronic kidney disease*	9	Rheumatoid arthritis*
10	Lupus*	11	Systemic heart disease**	12	Anti-osteoporosis use**
13	Heparin use**	14	Beta-blocker use**	15	Hypertension**
16	Deep vein thrombosis or pulmonary embolism**	17	Anticoagulant use**	18	Antidepression_TCA**
19	Antidepression_SSRI**	20	Hypercholesterolemia**	21	Statin use**
22	Family history of cardiovascular disease	23	Family history of cardiovascular disease before age 60	24	Heart failure*
25	Migraine*	26	Severe mental illness*	27	Vascular disease*
28	Atrial fibrillation*	29	on_bisphosbhonates_medication	30	Antipsychotic use**
31	Steroid use**	32	Erectile_dysfunction**	33	Cardiovascular disease
34	MI or Stroke	35	Established CVD*	36	Any fracture history
37	Hip fracture history	38	Shoulder fracture history	39	Spine fracture history
40	Wrist fracture history	41	BMI**	42	No. of GP visits**
43	No. of GP emergency visits**	44	eGFR**	45	SBP**
46	DBP**	47	No. of concomitant medicines**	48	Cholesterol measurement (HDL/LDL)**
49	std_SBP_recent_2	50	No. of previous fractures*	51	Age
52	Charlson score				

Figure 1. The list of patient characteristics; considered as candidate risk factors in this study. Abbreviations: * ever; ** in the year prior to index date; SES, socio-economic status; MI, myocardial infarction; BMI, body mass index; eGFR, estimated Glomerular Filtration Rate; SBP, cholesterol, systolic blood pressure; DBP, diastolic blood pressure; std_SBP_recent_2, Standard deviation of at least two most recent systolic blood pressure readings

E. Prediction Model development

To compare proposed prediction models with a reference model, we first fitted a logistic regression (LR) model to 21 QRISK-selected features (QRISK-LR). Next, an LR model with the least absolute shrinkage and selection operator (lasso) regularization technique (Lasso-LR) was developed using all 52 candidate predictors in Figure 1. Lasso-regularisation enabled data-driven feature selection before prediction. Finally, 4 ML models namely extreme gradient boosting (XGB), decision tree (DT), random forest (RF), and multi-layer perceptron neural network (NN) classifiers were applied to all 52 candidate features. Two versions of each ML model were prepared: a standalone intrinsic feature selection and prediction model, and another where a given ML model was used for feature selection alone and these features were then fitted to a subsequent LR model. All ML models were optimised using a hyperparameter grid search.

F. Evaluation and Performance Metrics

For internal validation, 10% of the dataset was randomly selected and held out as the test set; the remaining 90% was used as the training set.

Model performance was evaluated with regards to discrimination and calibration; the two criteria most widely used in literature for in silico evaluation of clinical risk prediction models [19-21]. Discrimination was measured using the area under the receiver operating characteristic (AUC) [26]; calibration was assessed by producing calibration plots of observed versus predicted probabilities and reporting the calibration slope (slope = 1 indicating perfect calibration) as well as the Brier score (score = 0 indicating perfect accuracy) [27].

III. Model Comparison and Results

A. Comparison of Models for Clinical Risk Prediction

In this section, we compared the performance of the four ML models with Lasso-LR and the reference QRISK-LR as defined in the previous section. All models were trained using the initial set of 52 features except for QRISK-LR. The Caret library version 6.0-92 in R version 4.2.1 was used to implement these models.

Fig. 2 presents the discrimination results. RF was the worst-performing model followed by DT with worse AUC than QRISK-LR (reference model). Whereas XGB and Lasso-LR had the highest AUCs, followed by NN.

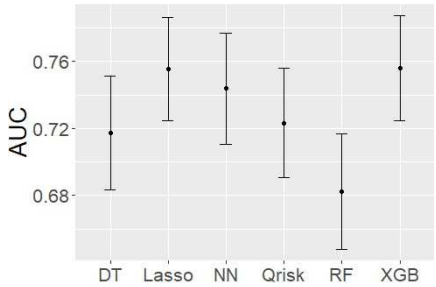

Figure 2. Comparison of discrimination among prediction models: x-axis depicts models as described in section E, with Lasso and Qrisk representing Lasso-LR and Qrisk-LR equivalents.

With regards to calibration performance (Table 1), RF and DT models had again the worst performance whilst the other models achieved similar calibration slopes and Brier scores.

TABLE 1. BRIER SCORE AND CALIBRATION SLOPES OF PREDICTION MODELS

Methods	Brier Score	Calibration Slope
Qrisk-LR	0.028	1.09
Lasso-LR	0.027	1.10
XGBoost	0.027	1.12
NeuralNet	0.027	1.09
Random Forest	0.029	0.45
Decision Tree	0.226	0.96

B. Comparison of Models for Risk Factor Identification

Assessment of feature importance can differ among ML models. To perform a comparative analysis of the risk factors selected as important features, the varImp function of the Caret library in R [28] was used. The library utilizes various methods

such as the absolute value of coefficients, Friedman's method [29], out-of-bag error estimation [30], and Gevrey's method [31] for selecting important features in LR, gradient boosting, random forest, and neural networks. To make importance measures comparable, varImp scales all measures of importance to a range of [0, 100]. We set a threshold of 10% for minimum importance and considered features above that as important or selected features. Figure 3 represents important features selected by each model. Patient characteristics selected as features or risk factors are presented on the horizontal axis as depicted in Figure 1, while the number of models that agree on the importance of each feature is shown on the vertical axis. The coloured blocks depict which ML models selected a given feature.

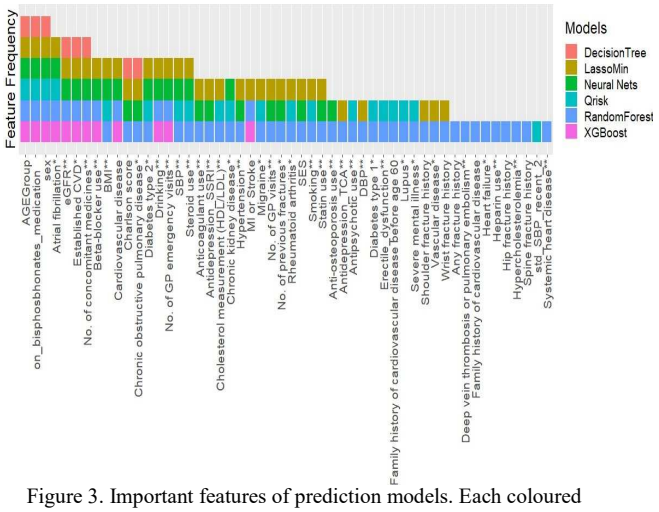

Figure 3. Important features of prediction models. Each coloured block presents the importance of the feature on the horizontal axis from the model's point of view specified with a colour.

The features age, sex, and bisphosphonate medication were considered important by all models. Moreover, 80% of prediction models agreed on the importance of the history of atrial fibrillation, one-year history of eGFR before OP diagnosis/fracture, established history of CVD, and one-year history of beta-blockers usage in predicting CVD death.

C. LR with Machine Learning Selected Risk Factors

In this section, we evaluated the performance of an LR model based on features which were selected by the ML models in the previous section, and we investigated whether LR could outperform ML models using the same set of selected features. In Figure 4, the different prediction model configurations are displayed on the horizontal axis, and their corresponding AUC values are shown on the vertical axis. Black lines present the AUC of LR as well as its confidence interval when we applied LR on feature sets presented in Figure 3. The blue lines show the discrimination of the standalone ML models.

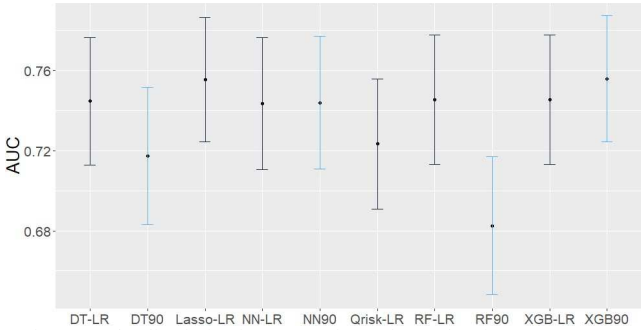

Figure 4. The discrimination comparison of LR when inputs are important features of prediction models is shown in black. The blue lines present the performance of ML models.

LR models fitted with features selected by DT and RF performed better than standalone DT and RF models. However, LR models fitted with features selected by NN and XGB did not perform better than standalone NN and XGB models in terms of AUC. Overall, with regards to discrimination and calibration, the best performing models were therefore XGB, NN, and Lasso-LR. All three outperformed the reference Qrisk-LR model which was based on a subset of expert-selected risk factors.

D. Effect of Sample Size

We finally investigated the effect of sample size on the performance of the three best models in the present study. To do so, we selected 20% of the initial dataset for testing purposes and trained three prediction models on small to large portions of the remaining dataset to examine the effect of training size on their performance. Specifically, the remaining 80% of the dataset was divided into eight different random subsets ranging in size from 10 to 80 percent of the training set. Models were trained on each subset, and their performance on the test set was compared. Figure 5 illustrates the discrimination of models for different training sizes. The two ML models achieved a similar AUC as Lasso-LR for the higher training sizes but underperformed with smaller training sizes, demonstrating ML models to be sensitive to training sample size variability, and Lasso-LR to be robust against smaller sample sizes.

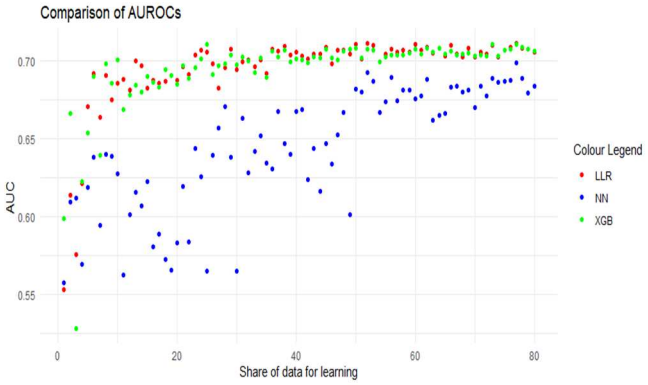

Figure 5. Discrimination comparison of LR, NN, and XGB prediction models with different train/test split proportions from 10 to 80% of the training sample size.

IV. Discussion and Conclusion

In this study, we compared machine learning models with existing clinical risk prediction tools for the estimation of

CVD mortality risk. Machine learning models were used for standalone feature selection and prediction, as well as just for feature selection or risk factor identification. Internal validation showed machine learning models XGB, Lasso-LR and NN to outperform the reference clinical model for the population of interest: patients with osteoporosis. It was shown that the performance of LR was improved when trained with ML-selected features. The sensitivity of ML models to small sample sizes was observed. Future work to explore the sample size conditions for superior model performance is ongoing.

These findings support the potential suitability of ML for data-driven risk factor identification in addition to clinical risk prediction. Further study into external validation, suitability against a variety of disease areas, and clinical explainability should be considered.

V. Ethical and Data Sharing Approval for Experiment

A. Ethical approval

The study was approved by the Independent Scientific Advisory Committee (protocol number 18_116R).

B. Data sharing

Data that supports the findings of this study was provided by the UK CPRD database. Availability of data is subject to protocol approval by CPRD's Research Data Governance Process.

References

[1] Johnston CB, Dagar M. Osteoporosis in older adults. Medical Clinics. 2020 Sep 1;104(5):873-84.

[2] Pineda-Moncusí, M., El-Hussein, L., Delmestri, A., Cooper, C., Moayyeri, A., Libanati, C., Toth, E., Prieto-Alhambra, D., & Khalid, S. (2022). Estimating the Incidence and Key Risk Factors of Cardiovascular Disease in Patients at High Risk of Imminent Fracture Using Routinely Collected Real-World Data From the UK. Journal of bone and mineral research: the official journal of the American Society for Bone and Mineral Research, 37(10), 1986–1996.

[3] Lampropoulos CE, Papaioannou I, D'cruz DP. Osteoporosis—a risk factor for cardiovascular disease?. Nature Reviews Rheumatology. 2012 Oct;8(10):587-98.

[4] Wilson PWF. Framingham Risk Score for Hard Coronary Heart Disease MDCalc [Available from: https://www.mdcalc.com/framingham-risk-score-hard-coronary-heart-disease], 2021.

[5] Andersson C, Johnson AD, Benjamin EJ, Levy D, Vasan RS. 70-year legacy of the Framingham Heart Study. Nature Reviews Cardiology. 2019 Nov;16(11):687-98.

[6] Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. Jama. 2001 Jun 13;285(22):2864-70.

[7] Gage BF, Van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode BS, Petersen P. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation. 2004 Oct 19;110(16):2287-92.

[8] Gage B. CHADS2 Score for Atrial Fibrillation Stroke Risk MDCalc [Available from: https://www.mdcalc.com/chads2-score-atrial-fibrillation-stroke-risk], 2021.

[9] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May 23;357.

[10] QRISK®3-2018 risk calculator. [Available from: https://www.qrisk.org/], 2018.

[11] Prieto-Alhambra D, Premaor MO, Fina Avilés F, Hermosilla E, Martinez-Laguna D, Carbonell-Abella C, Nogués X, Compston JE, Díez-Pérez A. The association between fracture and obesity is site-dependent: a population-based study in postmenopausal women. Journal of bone and mineral research. 2012 Feb;27(2):294-300.

[12] Premaor MO, Compston JE, Fina Avilés F, Pagès-Castellà A, Nogués X, Díez-Pérez A, Prieto-Alhambra D. The association between fracture site and obesity in men: a population-based cohort study. Journal of Bone and Mineral Research. 2013 Aug;28(8):1771-7.

[13] Prieto-Alhambra D, Premaor MO, Avilés FF, Castro AS, Javaid MK, Nogués X, Arden NK, Cooper C, Compston JE, Diez-Perez A. Relationship between mortality and BMI after fracture: A population-based study of men and women aged≥ 40 years. Journal of bone and mineral research. 2014 Aug;29(8):1737-44.

[14] Martinez-Laguna D, Tebe C, Javaid MK, Nogues X, Arden NK, Cooper C, Diez-Perez A, Prieto-Alhambra D. Incident type 2 diabetes and hip fracture risk: a population-based matched cohort study. Osteoporosis International. 2015 Feb;26(2):827-33.

[15] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. Ieee Access. 2017 Apr 26;5:8869-79.

[16] Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018 Oct;2(10):719-31.

[17] Sadek RM, Mohammed SA, Abunbehan AR, Ghattas AK, Badawi MR, Mortaja MN, Abu-Nasser BS, Abu-Naser SS. Parkinson's disease prediction using artificial neural network. Int. J. Academic Health Med. Res. 2019.

[18] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. InMachine learning for healthcare conference 2016 Dec 10 (pp. 301-318). PMLR.

[19] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P. Scalable and accurate deep learning with electronic health records. NPJ digital medicine. 2018 May 8;1(1):1-0.

[20] Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine. 2021 May 20;4(1):1-3.

[21] Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: transformer for electronic health records. Scientific reports. 2020 Apr 28;10(1):1-2.

[22] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Van Staa T, Smeeth L. Data resource profile: clinical practice research datalink (CPRD). International journal of epidemiology. 2015 Jun 1;44(3):827-36.

[23] NHS. Read Codes, Available at: https://digital.nhs.uk/services/terminology-and-classifications/read-codes (2019).

[24] WHO. ICD-10 online versions, Available at: https://icd.who.int/browse10/2016/e (2019)

[25] Rubin DB. Underlying Bayesian theory. Multiple imputation for nonresponse in surveys. 1987:1-76.

[26] Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006 Jun 1;27(8):861-74.

[27] Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020 Apr;27(4):621-33.

[28] Variable Importance, The caret Package, [Available from: https://topepo.github.io/caret/variable-importance.html], 2019.

[29] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001 Oct 1:1189-232.

[30] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

[31] Gevrey M, Dimopoulos I, Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling. 2003 Feb 15;160(3):249-64.