Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis

B.C.S. de Vries[1], J.H. Hegeman[1,2], W. Nijmeijer[1], J. Geerdink[3], C. Seifert[2], C.G.M. Groothuis-Oudshoorn[4]

[1] Department of Trauma Surgery, Ziekenhuisgroep Twente, Almelo, The Netherlands

[2] Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

[3] Department of Information & Organization, Ziekenhuisgroep Twente, Almelo, The Netherlands

[4] Health Technology and Services Research, Techmed Centre, University of Twente, Enschede, The Netherlands

Corresponding Author

Full name: Bram C.S. de Vries

Department: Department of Trauma Surgery

Hospital: Ziekenhuisgroep Twente

Street Name & Number: Zilvermeeuw 1

City, State, Postal code, Country: 7609 PP Almelo, The Netherlands

Tel: +31 88 708 7878

E-mail: bcsdevries@gmail.com

ORCID:

Bram de Vries: 0000-0002-0982-9873

Han Hegeman: 0000-0003-2188-2738

Wieke Nijmeijer: 0000-0002-5799-9690

Jeroen Geerdink: 0000-0001-6718-6653

Christin Seifert: 0000-0002-6776-3868

Karin Groothuis-Oudshoorn: 0000-0002-4875-5379

Mini-abstract

Four machine learning models were developed and compared to predict risk of a future Major Osteoporotic Fracture (MOF), defined as hip-, wrist-, spine- and humerus-fractures, in patients with a prior fracture. We developed a user-friendly tool for risk calculation of subsequent MOF in osteopenia patients, using the best performing model.

Abstract

Purpose:

Major osteoporotic fractures (MOFs), defined as hip-, wrist-, spine- and humerus-fractures, can have serious consequences regarding morbidity and mortality. Machine learning provides new opportunities for fracture prediction and may aid in targeting preventive interventions to patients at risk of MOF. Primary objective is to develop and compare several models, capable of predicting risk of MOF as a function of time in patients seen at the Fracture and Osteoporosis Outpatient Clinic (FO-clinic) after sustaining a fracture.

Methods:

Patients aged >50 years visiting an FO-clinic were included in this retrospective study. We compared discriminative ability (concordance-index) for predicting the risk on MOF with a Cox regression, Random Survival Forests (RSF) and an artificial neural network (ANN)-DeepSurv model. Missing data was imputed using multiple imputation by chained equations (MICE) or RSF's imputation function. Analyses were performed for the total cohort and a subset of osteopenia patients without vertebral fracture.

Results:

7578 patients were included, 805 (11%) patients sustained a subsequent MOF. Highest concordance-index in the total dataset was 0.697 (0.664-0.730) for Cox regression, no significant difference was determined between the models. In the osteopenia subset, Cox regression outperformed RSF (p=0.043 and p=0.023) and ANN-DeepSurv (p=0.043) with a c-index of 0.625 (0.562-0.689). Cox regression was used to develop a MOF risk calculator on this subset.

Conclusion:

We show that predicting risk of MOF in patients who already sustained a fracture can be done with adequate discriminative performance. We developed a user-friendly tool for risk calculation of subsequent MOF in patients with osteopenia.

Introduction

Osteoporosis is a major cause of fractures in elderly. Globally, 158 million individuals older than 50 years of age are estimated to be at high risk of osteoporotic fractures [1]. Consequently, 1 in 3 women and 1 in 5 men older than 50 years of age will suffer from an osteoporotic fracture [2]. The risk of sustaining an osteoporotic fracture is expected to have doubled in 2045 due to progressive ageing [1]. Major osteoporotic fractures (MOFs), defined as fractures of the hip, wrist, spine and humerus [3], have the highest incidence in osteoporotic patients and can have serious consequences. For example, hip fractures have a high rate of both morbidity and mortality [4]. Twenty to thirty-five percent of the patients admitted to the hospital with a hip fracture die within one year [5].

Patients with a prior fracture of any nature are known to be at risk for a future fracture. Warriner et al. (2011) described a relative risk of subsequent fracture at any site of 2.1 (1.6 – 2.7) compared to patients without a previous fracture, while the relative risk for a subsequent fracture after hip fracture, vertebral fracture or radius/ulna fracture is respectively 3.2 (2.3 – 4.5), 3.0 (2.0 – 4.3), and 2.7 (1.9 – 3.7) compared to patients without a previous fracture [6]. The current Dutch guideline 'Osteoporosis and fracture prevention' recommends bone mineral density (BMD) screening by means of a dual-energy X-ray absorptiometry (DXA scan) in patients older than 50 with a recent non-vertebral fracture [7]. Treatment decision for these patients is currently based on general classification of patients. Patients with a T-score < -2.5 and/or a vertebral fracture are treated with anti-osteoporotic medication, while other patients not meeting this criterion are not [7]. Recently, healthcare made a shift towards targeted prevention using more personalised medicine [8]. In the field of fracture prevention, risk assessment tools aim to personalize the prediction of future fracture risk to support treatment decisions. Risk identification for patients with osteopenia in absence of vertebral fracture is most relevant, as these patients are not standardly treated with anti-osteoporotic medication. Forty-eight fracture risk assessment tools were available in 2017. Three of those (FRAX®, Garvan and QFracture) have been tested in a population-based setting and focus on predicting direct fracture risk, with FRAX® being the most validated and used tool worldwide for hip fractures [9]. FRAX® uses twelve input parameters and is based on linear and non-linear combinations of risk-factors for future fracture [9]. However, it does not consider several important proven risk factors such as relevant comorbidities, patient's history and prior fall(s). Furthermore, it only incorporates the BMD of the hip, while measurement of BMD of the lumbar spine is more correlated to the risk of vertebral fractures [7]. The current guideline in the Netherlands does not advice to use FRAX® or any other fracture risk assessment tool for clinical decision making, but only in risk-communication to patients with osteopenia [7].

Recent advances in machine learning provide new opportunities for fracture prediction because of their ability to handle large numbers of input variables simultaneously. Especially, Artificial Neural Networks (ANNs) have shown promising results in different studies. For example, Tseng et al. (2013) designed an ANN outperforming standard logistic regression in the assessment of hip fracture risk in elderly patients [10]. Besides, Ho-le et al. (2017) showed that their ANN outperformed k-nearest-neighbour, support vector machines and logistic regression in the prediction of

hip fracture in post-menopausal women [11]. Artificial intelligence (AI) based models will potentially capture underlying trends and patterns, making predictions more accurate and therefore more useful in clinical practice [12].

The primary objective of this study is to develop and compare several traditional and non-traditional models, capable of predicting the risk of a subsequent MOF over time in patients who are seen at a Fracture and Osteoporosis Outpatient Clinic (FO-clinic) [13] after sustaining a fracture. The best performing model could serve as a risk-assessment tool in clinical practice for patients visiting an FO-clinic. To the best of our knowledge, this is the first study to design a model applicable to this specific population. Such risk-assessment tool may be helpful in targeting treatment to patients at high risk of MOF who do not currently meet the criteria for this treatment. Secondary aim of this study is to identify predictors of subsequent fractures in this population.

Methods

Study design

This retrospective cohort study was performed at the Ziekenhuisgroep Twente (ZGT) and was approved by the local ethics review committee. All patients aged >50 years who sustained a fracture and visited the FO-clinic of the ZGT between July 2011 and November 2019 were included. Primary endpoint of the study was the time to occurrence of a subsequent MOF, defined in line with Briot et al. (2013) as humerus-, wrist, clinical spine- or hip fracture [3]. Two separate datasets were created for analysis. The first dataset comprises all patients and will from now on be referred to as 'complete dataset'. The second dataset is a subset of the complete dataset. It contains only patients with osteopenia in the absence of a vertebral fracture as these patients are not standardly treated with anti-osteoporotic medication. This dataset will be referred to as the 'osteopenia subset'. All statistical procedures were performed for both datasets separately.

Data extraction

Data were extracted from the electronic health record database using structured query language (SQL) queries. Time till to occurrence of a MOF was extracted using diagnosis treatment combinations (DBCs in Dutch) as defined and labelled by the Dutch healthcare authority [14]. DBCs are healthcare products used for financial administration. The starting date for all patients was their visit to the FO-clinic. For patients with a DBC because of a MOF following their visit to the FO-clinic, the registration date of the DBC was used to calculate the time to occurrence of a MOF. Patients with no DBC of MOF were considered censored for the analysis, meaning that during the study time they had no subsequent MOF. Either the date of death or, when alive, the date of the end of the study (Nov. 2019) was used to calculate the follow-up time of censored patients. DBCs and completed forms from the FO-clinic were used to determine relevant comorbidities. DBCs do not provide information on the potential recovery of a comorbidity. Hence, we could not distinguish between history of a comorbidity and presence of a comorbidity. Therefore, the corresponding variables were combined into one single variable. Both clinical parameters and parameters regarding lifestyle were

extracted using FO-clinic forms. Biochemical parameters were extracted using a time window of 6 months prior to the visit at the FO-clinic until 1-week post visit.

Predictors

Demographic study parameters included age and gender of the patient. Relevant comorbidities selected are diabetes mellitus, cardiovascular diseases, inflammatory bowel disease, cerebral vascular accident, epilepsy, systemic auto-immune disease, rheumatoid arthritis, malabsorption disorder, renal insufficiency, collapse, delirium or dementia and vertigo. A history of fall(s), continuously being bedridden and a positive family history of a first-degree relative with either hip fracture or osteoporosis were collected for all patients. Number of children, use of combined oral contraceptive pill (COCP), breastfeeding to infants and duration of menopause (years) were collected for all female patients. Clinical parameters included use of corticosteroids, presence of vertebral fracture, reporting back pain, weight <67 kg, weight <60 kg, diminished height in recent years, height (cm), weight (kg) and moderate active hours per week. Radiographic variables collected were T-scores of the hip and lumbar spine. Parameters regarding lifestyle included dietary daily calcium intake (mg), >6 cups of coffee per day, frequent exposure to sunlight, diet includes fat fish (≥twice a week), vegetarian diet, use of vitamin supplements, and daily use of margarine. Biochemical parameters included erythrocyte sedimentation rate (ESR), plasma calcium, plasma albumin, plasma thyroid stimulating hormone (TSH), serum vitamin D3 and estimated glomerular filtration rate (eGFR[1]). Categorical variables are dichotomously described as yes/no, except for gender which is given as male/female.

Data preparation

Variables which only apply to a subset of the patients were transformed to two-part variables as described by Dziak and Henry (2017) [15]. This method was applied for the variables number of children, use COCP, breastfeeding to infants and duration of menopause as these were only relevant for female patients. Additionally, it was applied on eGFR as this variable contained both, data points defined as '>90' as well as numeric data points. Interactions terms were created after standardizing the continuous variables by mean centring. Unrealistic values were identified using boxplots, removed from the dataset and handled as missing data. Missing data was imputed using the MICE package in R [16], a total of 30 imputed datasets were used. More details are provided in Supplementary material 1.

Statistical analysis and modelling

Descriptive statistics are provided for both datasets and were compared between cases (occurrence of MOF) and controls (no occurrence of MOF) using the χ^2-square test for nominal variables and Mann-Whitney U test for continuous variables.

Our main outcome measure is the time to event, which is right censored survival data. We therefore used three survival modelling approaches to predict the risk of MOF as a function of time: Cox proportional hazard regression [17], Random Survival Forest (RSF) [18] and ANN Deepsurv [19].

---

[1] Determined with the CKD-EPI formula

The Cox proportional hazard regression is a traditional choice in survival modelling [17]. Cox regression assumes the hazards to remain constant over time (proportionality) and assumes linear effects of continuous variables on the log hazard function (linearity). Proportionality was checked using the Schoenfeld's residual correlation test [20]s. The linearity assumption was tested using the likelihood ratio test. Several transformations of continuous variables were considered and compared using the χ^2-statistic of the univariate Wald test [21]. We considered the following transformations for variables x: $\log(x)$, $\log(x)2$, $x2$, $\sqrt{x}$, and restricted cubic splines of x with 3 or 4 degrees of freedom. The coefficients of the variables transformed by a restricted cubic spline were calculated as demonstrated by Shepherd and Rebeiro (2018) using a reference category [22]. Variables for the final model were selected using Least Absolute Shrinkage and Selection Operator (LASSO), a modern variable selection technique which is preferable to stepwise selection methods [21]. It optimizes the bias-variance trade-off by determining an optimum for the penalty factor (λ) using cross-validation. In line with Steyerberg (2011), the penalty factor λ (or λ + 1 standard error) was determined using tenfold cross-validation and subsequently averaged [21].

RSF, developed by Ishwaran et al (2008), uses multiple decision trees for predictive modelling and is designed for right-censored survival data [18]. We implemented two RSF models. The first model was trained and validated on every single dataset created by the MICE imputation, from now on referred to as RSF-MICE. The second model used the imputation method of the RSF algorithm itself and will be referred to as RSF-regular. The hyperparameters used for RSF can be found in Supplementary material 2. RSF can give insight in the prediction models and its variates by calculating variable importance of every variable [23]. The variable importance of every specific variable is determined by analysing out-of-bag samples, detailed information is given in Supplementary material 1. Due to incompatibility problems with the pool function of MICE, a single imputed dataset was used to determine variable importance for RSF-MICE.

Deepsurv is a Cox proportional hazards deep neural network designed by Katzman et al (2018) [19]. We used Python to implement the DeepSurv model (ANN-DeepSurv). Data was standardized prior to model development. To identify the best hyperparameters [24], we applied grid search with tenfold cross-validation on a single imputed dataset. Constraints for the hyperparameters were set manually, weighing both computational time and model performance. We used the Adaptive Moment Estimation (Adam) optimizer [25] for training ANN-Deepsurv and left the L2-regularization to the default value (1e-4). L2 regularization is frequently used in ANN's to reduce model complexity during training and reduce overfitting. It can be viewed as an implicit (latent) feature selection method [26]. Hyperparameters and their constraints can be found in Supplementary material 3. Hyperparameters of the model with the largest c-index were selected for final evaluation. As the ANN-DeepSurv model showed large variance in the c-index, 10 final models were constructed per dataset and subsequently averaged.

Evaluation of model performance

Discriminative performance of the four models was evaluated using the concordance index (c-index), the most commonly used performance measure to indicate discriminative ability for survival models [21]. Like AUC, a c-index of 0.5 is comparable to random guessing while a c-index of 1.0 reflects perfect discrimination. Tenfold cross-validation was used to assess the c-index and its confidence interval. The c-indexes were tested for statistically significant differences using two-sample Mann-Whitney U tests (two-sided). To contextualise the predictive accuracy of the best performing models, we additionally constructed receiver operating curves at different points in time.

Summary of methods

Four different models (Cox regression, RSF-MICE, RSF-regular and ANN-DeepSurv) were constructed and their c-indexes were compared. All models used a form of feature selection, thereby reducing complexity of the models and returning a feasible number of risk factors for use in clinical practice. Missing data points were imputed using MICE before applying Cox regression, RSF-MICE and ANN-DeepSurv, while RSF-regular used its integrated imputation method. All models were trained and tested on two datasets: the complete dataset and the osteopenia subset. The best performing model in the osteopenia subset was used to construct a risk calculator predicting the three- and five-year risk of MOF. This risk calculator was designed using Shiny, a R package to implement interactive web applications.

P-values <0.05 are considered statistically significant. Software used for data preparation and analysis in this research were R (R Core Team 2019, version 3.6.1) [16] and Python 3.7 (Scikitlearn and PySurvival) [27, 28].

Results

In this study 7578 patients were included, 5014 (74%) were female. A total of 805 (11%) patients sustained a subsequent MOF. Median time to sustaining an MOF was 114 weeks (Interquartile range (IQR) = 153), while the median follow-up time of patients with no MOF was 192 weeks (IQR = 224). Median age for all patients was 68 years (IQR = 17), and 74 (IQR = 15) and 67 (IQR = 16) for patients who sustained a subsequent MOF and censored patients, respectively. The osteopenia subset consists of 1770 patients of which 1367 (77%) were female. In this subset, 165 (9%) patients sustained a subsequent MOF. Median time to sustaining an MOF was 118 weeks (IQR = 147), while the median follow-up time of patients with no MOF was 159 weeks (IQR = 217). Median age for all patients in the osteopenia subset was 67 years (IQR = 17), and 72 (IQR = 15) and 67 (IQR=16) for patients who sustained a subsequent MOF and censored patients, respectively.

Overall, 2271 (30%) cases had missing data, while for the osteopenia subset 590 (33%) cases were incomplete. The percentage of missing values across all variables varied between 0% and 23% (for history of fall(s)). The primary endpoints, occurrence of a MOF and time to occurrence of a MOF, were complete in both datasets. Significant correlation with MOF in univariate analysis was found in various covariates in both datasets. Detailed information on descriptive statistics and missing data of each variable are shown in Table 1.

Table 1: Descriptive statistics for both the complete dataset (left) and the osteopenia dataset (right)

	Missing	No MOF (n=6773)	MOF (n=805)	p-value	Missing	No MOF (n=1605)	MOF (n=165)	p-value
>6 cups of coffee per day	5,2%	732 (11.4)	58 (7.7)	0.002	3,5%	166 (10.7)	15 (9.6)	0.757
Age (years)	0,0%	67 [60, 76]	74 [66, 81]	<0.001	0,0%	67 [59, 75]	72 [65, 80]	<0.001
Breastfeeding to infants	7,5%	2725 (43.4)	395 (53.4)	<0.001	6,8%	688 (46.0)	74 (48.1)	0.692
Cardiovascular disease	0,0%	2560 (37.8)	347 (43.1)	0.004	0,0%	590 (36.8)	74 (44.8)	0.05
Cerebral vascular accident	0,0%	733 (10.8)	106 (13.2)	0.052	0,0%	165 (10.3)	17 (10.3)	1.000
Daily use of margarine	3,8%	5979 (91.7)	709 (92.4)	0.511	3,2%	1416 (91.0)	142 (89.9)	0.745
Decreased renal function	1,7%	4741 (71.2)	567 (71.8)	0.772	1,1%	1128 (71.1)	116 (71.2)	1.000
Delirium or dementia	0,0%	202 (3.0)	46 (5.7)	<0.001	0,0%	45 (2.8)	12 (7.3)	0.004
Diabetes Mellitus	0,0%	837 (12.4)	119 (14.8)	0.057	0,0%	167 (10.4)	24 (14.5)	0.133
Diet includes fat fish	6,8%	2376 (37.6)	248 (33.5)	0.031	5,8%	568 (37.5)	62 (40.8)	0.473
Dietary daily calcium intake (milligram)	0,5%	865 [625, 1015]	845 [655, 1015]	0.804	0,2%	845 [625, 985]	810 [630, 970]	0.664
Diminished length in recent years	8,6%	3498 (56.5)	526 (71.0)	<0.001	7,4%	782 (52.6)	107 (70.4)	<0.001
Duration of menopause (years)	11,8%	13 [0, 24]	21 [9, 30]	<0.001	11,1%	13 [0, 24]	19 [4, 29]	0.001
eGFR	1,7%	79 [67, 90]	76 [63, 90]	0.001	1,1%	79 [68, 90]	77 [62, 90]	0.243
Epilepsy	0,0%	122 (1.8)	37 (4.6)	<0.001	0,0%	29 (1.8)	6 (3.6)	0.189
Erythrocyte sedimentation rate	3,1%	9 [5, 18]	11 [5, 21]	<0.001	2,1%	9 [5, 16]	8 [3, 15.50]	0.296
Ever being bedridden	2,6%	618 (9.4)	83 (10.7)	0.249	1,8%	152 (9.7)	17 (10.4)	0.857
Frequent exposure to sunlight	2,7%	5918 (89.7)	633 (82.2)	<0.001	1,9%	1432 (90.9)	140 (87.0)	0.134
Gender (female)	0,0%	5014 (74.0)	673 (83.6)	<0.001	0,0%	1237 (77.1)	130 (78.8)	0.687
History of collapse	0,0%	145 (2.1)	18 (2.2)	0.962	0,0%	40 (2.5)	5 (3.0)	0.874
History of fall(s)	22,7%	1301 (24.8)	197 (32.2)	<0.001	21,4%	293 (23.1)	40 (32.5)	0.026
Inflammatory bowel disease	0,0%	79 (1.2)	7 (0.9)	0.565	0,0%	20 (1.2)	2 (1.2)	1.000
Length (cm)	15,0%	168 [162, 174]	165 [160, 170.50]	<0.001	16,7%	168 [162, 173]	166 [161, 171]	0.052
Malabsorption disorder	0,0%	31 (0.5)	5 (0.6)	0.714	0,0%	12 (0.7)	1 (0.6)	1.000
Moderate active hours per week	2,8%	21 [14, 40]	21 [14, 40]	0.009	2,0%	21 [14, 40]	21 [14, 40]	0.514
Number of children	7,6%	2 [0, 3]	2 [1, 3]	<0.001	7,5%	2 [0, 3]	2 [0, 3]	0.155
Plasma albumin	3,3%	38 [36, 40]	38 [35.50, 40]	<0.001	2,2%	38 [36, 40]	38 [36, 40]	0.878
Plasma calcium	2,9%	2.38 [2.32, 2.45]	2.37 [2.31, 2.45]	0.017	2,1%	2.39 [2.33, 2.45]	2.39 [2.31, 2.46]	0.721
Plasma TSH	2,5%	1.80 [1.20, 2.60]	1.80 [1.11, 2.70]	0.260	1,6%	1.80 [1.20, 2.60]	1.80 [1.10, 2.70]	0.938
Positive family history	1,9%	1631 (24.6)	201 (25.3)	0.679	1,9%	429 (27.3)	39 (23.8)	0.384
Presence of vertebral fracture	1,6%	1264 (19.0)	227 (28.5)	<0.001	n.a.	n.a	n.a.	n.a.
Renal insufficiency	0,0%	129 (1.9)	21 (2.6)	0.222	0,0%	32 (2.0)	4 (2.4)	0.934
Reporting back pain	8,5%	2888 (46.5)	378 (51.9)	0.006	8,5%	526 (35.8)	57 (38.0)	0.657
Rheumatoid arthritis	0,0%	249 (3.7)	42 (5.2)	0.040	0,0%	54 (3.4)	9 (5.5)	0.246
Serum vitamin D$_3$	9,7%	52 [36, 67]	47 [30, 65]	<0.001	7,2%	53 [37, 68]	46 [32, 68]	0.082
Systemic autoimmune disease	0,0%	268 (4.0)	26 (3.2)	0.361	0,0%	64 (4.0)	2 (1.2)	0.115
T-score hip	3,8%	-1.30 [-1.90, -0.60]	-1.50 [-2.10, -0.90]	<0.001	0,0%	-1.60 [-1.90, -1.20]	-1.70 [-2, -1.30]	0.019
T-score lumbar spine	1,7%	-1.50 [-2.30, -0.70]	-1.70 [-2.50, -0.92]	<0.001	0,0%	-1.80 [-2.20, -1.40]	-1.80 [-2.10, -1.50]	0.497
Use of COCP	9,1%	2830 (45.8)	295 (41.6)	0.038	7,9%	736 (49.6)	63 (42.6)	0.121
Use of corticosteroids	4,2%	415 (6.4)	55 (7.3)	0.364	3,8%	101 (6.5)	11 (6.9)	0.99
Use of vitamin supplements	4,2%	2789 (42.9)	307 (40.8)	0.289	3,1%	713 (45.9)	62 (38.5)	0.089
Vegetarian diet	5,2%	159 (2.5)	19 (2.5)	1.000	4,4%	46 (3.0)	3 (1.9)	0.592
Vertigo	0,0%	125 (1.8)	14 (1.7)	0.941	0,0%	35 (2.2)	1 (0.6)	0.282
Weight (kg)	13,8%	73.35 [65, 84]	70 [62, 80]	<0.001	15,8%	70.60 [63, 80]	70.50 [64, 78]	0.758
Weight <60 kg	2,5%	797 (12.1)	137 (17.7)	<0.001	2,0%	205 (13.0)	18 (11.2)	0.586
Weight <67 kg	2,7%	1845 (28.0)	273 (35.1)	<0.001	2,1%	526 (33.5)	52 (31.9)	0.745

Abbreviations: COCP = Combined Oral Contraceptive Pill, CI = Confidence Interval, eGFR = estimated Glomerular Filtration Rate. Categorical variables are described as number (percentage) while continuous variables are described as median [1st quartile – 3rd quartile].

The overall survival rate with MOF as event at year 1, year 3 and year 5 were respectively 0.967 (0.963-0.971), 0.914 (0.907-0.922), 0.866 (0.857-0.876) in the complete dataset. For the osteopenia subset, this rate was determined as 0.974 (0.966-0.982) for year 1, 0.922 (0.908-0.937) for year 3 and 0.865 (0.844-0.887) for year 5.

Cox regression

To meet the linearity assumption, the variables number of children and moderate active hours per week were transformed by the square root function, while ESR was log squared. Continuous variables duration of menopause, dietary daily calcium intake, plasma TSH and serum vitamin D were transformed to linear using restricted cubic splines with 3 degrees of freedom. Plasma calcium was transformed using restricted cubic splines with 4 degrees of freedom. Cox proportional hazard assumptions were met: global χ^2 was 53.3 (p-value 0.777) for the complete dataset and global χ^2 was 59.9 (p-value 0.516) for the osteopenia subset implying no deviation of the proportional hazard assumption. The penalty factor λ of LASSO was determined to be log(-4.14) and log(-4.41) for the complete dataset and osteopenia subset, respectively.

Age, prior fall(s), current vertebral fracture, history of epilepsy and duration of menopause were all independently associated with occurrence of MOF in the complete dataset. Hazard ratio of these variables ranged from 1.010 to 2.159. Moreover, interaction of age and T-score of the hip was also independently associated with this primary outcome measure with a hazard ratio of 1.010 (1.003 – 1.017). The variables frequent exposure to sunlight and T-score of the hip showed a reduction of risk for MOF, with a hazard ratio of 0.731 (0.602-0.888) and 0.386 (0.233-0.639), respectively. Both 0 and 50 years of menopause showed a reduction of risk for MOF (hazard ratio of 0.875 (0.836-0.917) and 0.853 (0.779-0.994) compared to 10 years, respectively). Results of the Cox regression in the complete dataset are shown in Table 2.

For patients in the osteopenia subset, prior fall(s), change in length in recent years, fat fish diet and a decreased renal function were significantly correlated with occurrence of MOF. The hazard ratio increased for diminishing renal function to 3.268 (2.672-3.997) for an eGFR of 10 mL/min/1.73m^2 compared to 70 mL/min/1.73m^2. Results of the Cox regression of the osteopenia subset are shown in Table 2. The hazard function of the variables transformed using restricted cubic splines are plotted in Supplementary material 3.

The Cox regression model returned a tenfold cross validated c-index of 0.697 (0.664 – 0.730) for the complete dataset and 0.625 (0.562 – 0.689) for the osteopenia subset.

Table 2: Results of cox proportional hazard regression of a) the complete dataset and b) the osteopenia subset

a) Complete dataset

Variable		Hazard Ratio (CI)	p-value
Age		1.052 (1.035-1.069)	<0.001
Gender (female)		1.329 (0.751-2.353)	0.329
History of fall(s)		1.357 (1.128-1.631)	0.001
Presence of vertebral fracture		1.425 (1.215-1.671)	<0.001
Epilepsy		2.159 (1.545-3.018)	<0.001
Frequent exposure to sunlight		0.731 (0.602-0.888)	0.002
T-score hip		0.386 (0.233-0.639)	<0.001
Duration of menopause (years)			<0.001
	0	0.875 (0.836-0.917)	
	10*	1.000	
	20	1.075 (1.040-1.110)	
	30	1.034 (0.986-1.083)	
	40	0.941 (0.881-1.004)	
	50	0.853 (0.779-0.994)	
Interaction of age and T-score hip		1.010 (1.003-1.017)	0.004

* used as reference category

b) Osteopenia subset

Variable		Hazard Ratio (CI)	p-value
Age		1.025 (0.963-1.092)	0.430
Gender (female)		0.907 (0.615-1.337)	0.620
History of fall(s)		1.577 (1.060-2.347)	0.025
Cardiovascular disease		1.233 (0.885-1.718)	0.216
Delirium or dementia		1.544 (0.823-2.895)	0.176
Diminished length in recent years		1.558 (1.073-2.262)	0.020
Moderate active hours per week		0.997 (0.986-1.008)	0.619
Diet includes fat fish		1.495 (1.069-2.090)	0.019
Decreased renal function		3.218 (1.496-6.293)	0.003
eGFR			<0.001
	70*	1.000	
	50	1.343 (1.270-1.421)	
	30	2.095 (1.843-2.382)	
	10	3.268 (2.672-3.997)	
T-score hip		0.828 (0.059-11.56)	0.889
Interaction of age and T-score hip		0.999 (0.964-1.035)	0.943

* used as reference category

RSF – MICE

Cross-validation (tenfold) returned a c-index of 0.688 (0.652-0.723) and 0.594 (0.536-0.651) for the complete dataset and osteopenia subset respectively. Age (0.022), T-score of the hip (0.014) and duration of the menopause (0.013) showed highest variable importance in the complete dataset. In the osteopenia dataset, age (0.020), duration of menopause (0.007) and the comorbidity delirium or dementia returned (0.006) the highest variable importance. The fifteen variables with largest variable importance are plotted in Figure 1 for both datasets.

RSF – regular

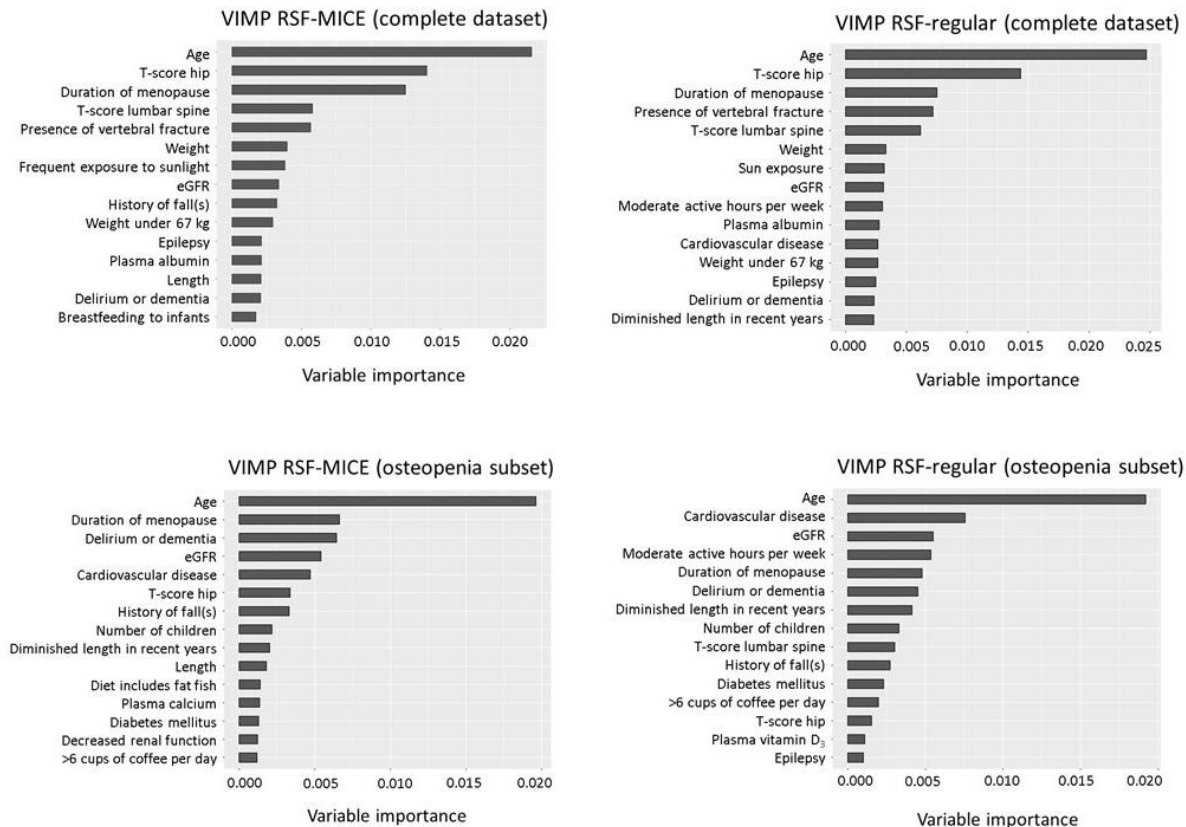A c-index of 0.687 (0.679-0.695) for the complete dataset and 0.593 (0.577 – 0.608) for the osteopenia subset was determined with tenfold cross-validation. Most important variables in the complete dataset were age (0.024), T-score of the hip (0.013) and current vertebral fracture (0.007). Age (0.019), cardiovascular disease (0.008) and eGFR (0.005) showed highest variable importance in the osteopenia subset. The fifteen variables with largest variable importance are plotted in Figure 1 for both datasets.

Figure 1: Variable importance (VIMP) of the fifteen variables with highest variable importance for every RSF model

ANN-Deepsurv

A c-index of 0.670 (0.592 – 0.747) for the complete dataset and 0.588 (0.506 – 0.671) for the osteopenia subset were determined using tenfold cross-validation. Results of the optimal hyperparameters returned by grid-search are shown in Supplementary material 4 for both datasets.

Model comparison

For the complete dataset, no significant difference was found between discriminative ability of the models. In the osteopenia subset, the Cox regression model significantly outperformed the RSF-MICE model (p=0.043), the RSF-regular model (p=0.023) and the ANN-DeepSurv model (p=0.043) on discriminative ability. Comparison of the c-

indexes of the four models are given in Figure 2 for each dataset. Receiver operating curves with corresponding area under the curve of the Cox regression models at different time points are given in figure 3.
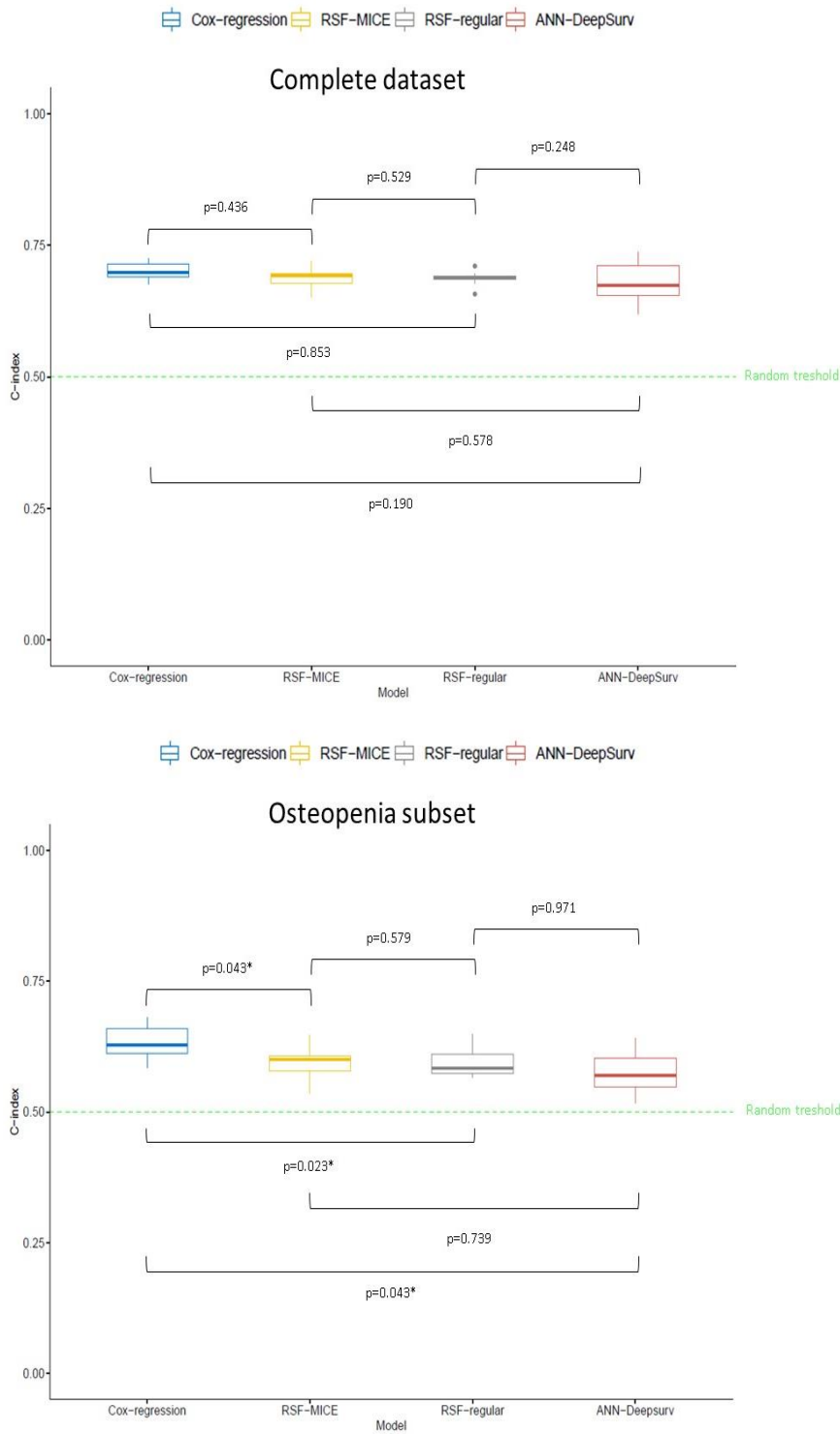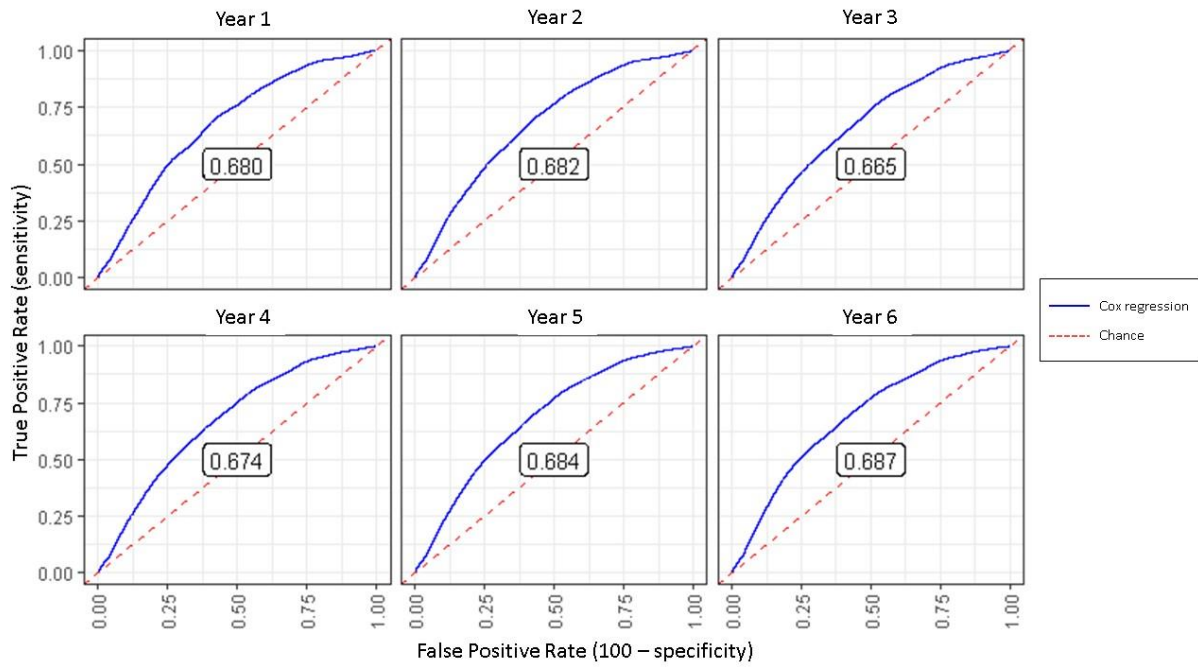

Figure 2: Boxplot of the c-index of the predictive models: Cox regression, RSF-MICE, RSF-regular and ANN-DeepSurv. Corresponding p-value of the unpaired two-sample Mann Whitney U test for every individual comparison is shown. Sign * indicates statistical significance (p<0.05).
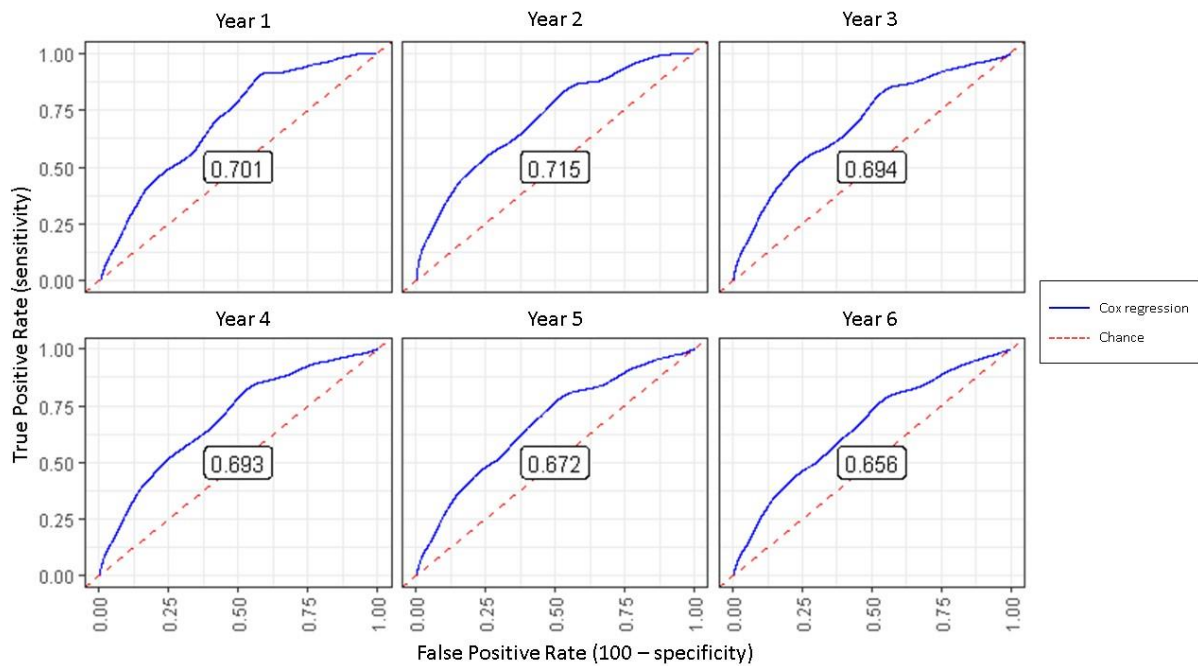
Figure 3: Receiver operating curves at different time points for the Cox regression models. The corresponding area under the curve at that specific time point is given in the centre of each plot.

Major Osteoporotic Fracture Risk Calculator

In the osteopenia subset, Cox regression model was used to develop a risk calculator, giving the 3- and 5-year risk of a MOF as an output. To construct the major osteoporotic risk calculator, we used the hazard ratio of the variables and the baseline function of the cox regression model of the osteopenia subset. An example of this risk calculator is shown in Figure 4.

Figure 4: Major Osteoporotic Fracture Risk Calculator

Discussion

This retrospective pilot study is, to the best of our knowledge, the first study to compare both traditional and non-traditional models capable of predicting the risk of sustaining a subsequent MOF in patients who were seen at the FO-clinic after already sustaining a fracture. We developed four models that adequately predict the risk of MOF over time in these patients and determined the predictive ability of Cox regression model, two RSF models and ANN-DeepSurv model for patients at the FO-clinic to be comparable. The discriminative ability of all models in the osteopenia subset is found to be lower compared to the total dataset, with the Cox regression model outperforming the RSF and ANN-DeepSurv models in osteopenia dataset. Finally, we designed a MOF risk calculator for fracture patients with osteopenia presenting themselves at a FO-clinic.

The non-traditional models ANN-DeepSurv (c-index 0.670, CI: 0.592 – 0.747) and RSF (c-index 0.687, CI: 0.679-0.695 / c-index: 0.688, CI: 0.652-0.723) did not outperform Cox regression (c-index: 0.697, CI: 0.664 – 0.730) in our complete dataset and returned significantly lower c-indexes in the osteopenia subset. This in contrast to the original DeepSurv study of Katzman et al. (2018), which showed a higher c-index on real life datasets for both RSF and

DeepSurv [19]. Likewise, Kim et al. (2019) showed that RSF and DeepSurv outperformed Cox regression in survival prediction of oral cancer patients [29]. A possible explanation for this contrary finding might be the number of variables used in this research. We used 46 and 45 variables for the complete dataset and osteopenia subset respectively, while aforementioned studies used 5 to 14 variables. Besides, our dataset (event rate = 11%) was more imbalanced compared to the datasets of these studies, with event rates ranging from 17% to 68%. Furthermore, we used models in the context of fracture prediction, while Katzman et al. (2018) predicted survival in the fields of cardiology and oncology. These factors might have resulted in lower discriminative ability of these models. Another explanation might be the differences in development of the Cox regression model. In our study, we used both LASSO and restricted cubic splines to optimize the bias-variance trade-off in the Cox regression model and met its assumption of linearity. In this way we were able to include non-linear effects into the Cox regression model. Katzman et al. (2018) and Kim et al. (2019) did not specify if they optimized the Cox regression model, nor if they transformed variables to meet the linearity assumptions [19, 29]. Therefore, the Cox regression models in their studies may not perform optimally. Cox regression outperforming the other models in the osteopenia subset may be caused by the lower sample size in this dataset compared to the complete dataset, as both RSF and ANN-DeepSurv showed more overfitting in the osteopenia subset compared to the complete dataset.

Literature comparing machine learning principles with traditional statistics for prediction of MOF is scarce. Forgetta (2018) and Nissinen (2019) both used machine learning models to predict osteoporotic fractures, but limited themselves to use of genotypes and DXA imaging respectively [30, 31]. Kruse (2017) and Tseng (2013) used multiple sources of information but focused solely on hip fracture. However, they both concluded that machine learning techniques can outperform traditional statistics in hip fracture prediction [10, 32]. Standard logistic regression analysis, although not able to handle censored data, is more often used to assess risk of MOF. Briot et al. (2013) analysed the predictive ability of the FRAX® tool for MOF over 6 years in postmenopausal women [3], while Ensrud et al. (2009) compared the FRAX®-tool to the use of BMD and age alone for 10-years of follow-up [33]. The incidence rate of 11% of MOF in our study was relatively high compared to the 4.9% of Briot et al. (2013), most likely due to the selection of our population at the FO-clinic. All patients who visited this clinic sustained a recent fracture and are therefore known to be at risk of a new fracture [34]. Ensrud et al. (2009) reported an even higher incidence of 16.6% but considered a longer follow-up period. Briot et al. (2013) returned a c-index of 0.69 (0.63-0.75) for the FRAX®-tool, while Ensrud et al. (2009) report a likewise c-index of 0.69 (0.67-0.70) for their model. These results are comparable to our models, with a maximum c-index of 0.70 (0.66 – 0.73) [3]. Reber et al. (2018) used survival analysis by the means of a Cox proportional hazard regression model to develop a fracture risk assessment tool based on claims data. They report a low MOF incidence (2.6%), probably due to their short follow-up period of 2 years. They determined a c-statistic of 0.70 (0.69-0.71), but also noted a decrease in c-statistic for 5 year follow-up [35]. Again, this is comparable to our results. The main difference, however, is that our model is applicable to a selected population of the Fracture and Osteoporosis outpatient clinic where all patients have a T-score < -1.0 or a vertebral fracture and sustained a fracture. The FRAX®-tool and the other studies considered a more general population.

Most risk factors for MOF as identified in this study, such as history of fall(s) or presence of vertebral fracture, are well-known risk factors for future fracture [3, 36–38]. An explanation for the increased hazard ratio of patients with epilepsy is given by Zhao et al. (2015). They stated that these patients are at risk mostly due to (myoclonic) seizures [39]. Dietary use of fat fish as a risk factor for subsequent fracture, however, is not in line with current literature. Perna et al. (2017) performed a systematic review and concluded that a fish dietary pattern has no negative effect on bone quality [40]. The negative effect of dietary use of fat fish on sustaining a MOF in this study may be caused by specific advice given to patients prior to sustaining the fracture which led to a visit to the FO-clinic. Reasons for this advice could be a history of fractures, a diminished T-score or a lower serum vitamin D in order to increase the vitamin D level. As these patients may have received this advice in the years prior to inclusion in this study, this effect may therefore mimic a lower bone quality. However, this could not be verified as no data regarding this issue was available.

At a FO-clinic, risk identification for patients with osteopenia in absence of vertebral fracture is most relevant, as these patients are not standardly treated with anti-osteoporotic medication. Anti-osteoporotic medication, such as bisphosphonates, is well-known to prevent osteoporotic fractures with relative risk reduction ranging from 45% to 53% [41]. In our study, 165 osteopenia patients (9%) sustained a subsequent MOF. As these patients could possibly benefit from anti-osteoporotic medication, identification of these patients is important. To the best of our knowledge, this is the first study to develop and compare several models capable of predicting the risk of a subsequent MOF for this specific group. The discriminative ability of our models in this population was lower compared to the total population at the FO-clinic. Less distinctive patient characteristics could explain this finding, as patients were pre-selected on their T-scores and absence of vertebral fracture. Both variables were important for the predictive models of the complete dataset, as reflected by their hazard ratio and variable importance. This study is a first step in the development of models predicting the risk of a subsequent MOF for patients with osteopenia who already sustained a fracture. We translated the best performing model to a user-friendly calculator for 3- and 5-year risk of a MOF. If further refined and both prospectively and externally confirmed in larger datasets, this risk calculator might aid in the identification of patients at risk of subsequent fracture in this population and therefore help targeting treatment to patients at highest need.

This study includes several limitations, which should be acknowledged. At first, several important variables, such as the number of falls in the last year, were not available due to the retrospective design of the studies. Both vertebral fracture assessment (VFA) and intoxications including smoking and alcohol use were not uniformly extractable from the electronic health record and could therefore not be considered. When clearly registered, several important variables including VFA might be added in future (prospective) research and may increase the discriminative ability of the models. Besides, the treatment decision for every patient was registered using free text and therefore not suitable for analysis. We therefore assumed that, in line with the protocol in our hospital, all patients with a T-score < -2.5 or a vertebral fracture were treated with anti-osteoporotic medication while all patients with osteopenia were not. However, individual circumstances may have led to treatment in osteopenia patients and thereby our results may be troubled.

Future research, with a categorical registration of the initiated therapy, is needed to confirm and improve our results. Natural language processing may be a future solution to determine treatment choice retrospectively.

Secondly, we experienced serious technical constraints of different software and packages in combination with survival data. Especially the ANN-DeepSurv model may have suffered from this constraint, as the hyperparameter search was performed on a single dataset. Technical development of functions for survival data is required to be able to deal with these problems properly in future research.

Lastly, using financial information of DBCs of our hospital has some drawbacks. Although most patients are loyal to our hospital, patients on the border of our catchment area might have been treated for a MOF in a nearby hospital and therefore incorrectly been labelled as event-free. Besides, using DBCs may result in a bias towards severe comorbidities, as patients with mild comorbidities are treated by the general practitioner (GP). GPs in the Netherlands do not make use of DBCs. Hence, mild comorbidities were not registered. Financial data on the other hand is verified by several institutions and might be more reliable than questionnaires, as the latter depends on patients' memory in an aging population. We suggest that results of the comorbidities are interpreted with caution and acknowledge the need of verification in future research.

In our study, the traditional model (Cox regression) outperformed the non-traditional models (RSF and ANN-DeepSurv) in the osteopenia subset, while no difference was found in the complete dataset. However, we need to consider that the Cox regression model was combined with new techniques like LASSO and restricted cubic splines. Nowadays, more and more innovations are used in traditional statistics, thereby blurring the line with non-traditional models such as RSF's and ANN's. As both RSF and ANN-DeepSurv showed more overfitting compared to the Cox regression model in the osteopenia subset, we recommend future studies with more data when applying non-traditional models as these techniques are promising. Our advice is to include more cases and use data of different hospitals. For future studies with more cases, we also recommend dividing the osteopenia patients in three tertiles based on their T-score. As relatively more events may occur in the group with the lowest T-score, a model based on this subgroup may return a higher predictive ability. A preliminary analysis in our dataset using three tertiles in the osteopenia subset showed promising results.

In conclusion, this study is the first study to compare both traditional and non-traditional models capable of predicting the risk of sustaining a subsequent MOF in patients who were seen at the FO-clinic after already sustaining a fracture. We show adequate and comparable discriminative performance of a Cox regression model, RSF models and ANN-DeepSurv model in a large population of a FO-clinic. In patients with osteopenia, Cox regression outperformed both RSF and ANN-DeepSurv and we developed a user-friendly tool for risk calculation of a subsequent MOF within 3- and 5-years. Further research, with a clear registration of important variables, is recommended to refine and validate this risk calculator and confirm our results. Although we acknowledge several limitations in our research, this study may be the starting point for models which identify patients with osteopenia at significant risk of subsequent fracture and therefore in need for a patient tailored treatment strategy.

Statements

References

1. Odén A, McCloskey E V., Kanis JA, et al (2015) Burden of high fracture probability worldwide: secular increases 2010–2040. Osteoporos Int 26:2243–2248. https://doi.org/10.1007/s00198-015-3154-6
2. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 17:1726–1733. https://doi.org/10.1007/s00198-006-0172-4
3. Briot K, Paternotte S, Kolta S, et al (2013) FRAX®: Prediction of Major Osteoporotic Fractures in Women from the General Population: The OPUS Study. PLoS One 8:e83436. https://doi.org/10.1371/journal.pone.0083436
4. Jürisson M, Pisarev H, Kanis J, et al (2016) Quality of life, resource use, and costs related to hip fracture in Estonia. Osteoporos Int 27:2555–2566. https://doi.org/10.1007/s00198-016-3544-4
5. Goldacre MJ, Roberts SE, Yeates D (2002) Mortality after admission to hospital with fractured neck of femur: database study. BMJ 325:868–9. https://doi.org/10.1136/bmj.325.7369.868
6. Warriner AH, Patkar NM, Yun H, Delzell E (2011) Minor, Major, Low-Trauma, and High-Trauma Fractures: What Are the Subsequent Fracture Risks and How Do They Vary? Curr Osteoporos Rep 9:122–128. https://doi.org/10.1007/s11914-011-0064-1
7. CBO Richtlijn Osteoporose en Fractuurpreventie-2011. https://www.volksgezondheidenzorg.info/bestanden/documenten/cbo-richtlijn-osteoporose-en-fractuurpreventie-2011. Accessed 21 Oct 2019
8. Weda M, Jansen ME, Vonk RAA (2017) Personalised medicine - Implementatie in de praktijk en data-infrastructuren
9. El-Hajj Fuleihan G, Chakhtoura M, Cauley JA, Chamoun N (2017) Worldwide Fracture Prediction. J Clin Densitom 20:397–424. https://doi.org/10.1016/j.jocd.2017.06.008
10. Tseng WJ, Hung LW, Shieh JS, et al (2013) Hip fracture risk assessment: Artificial neural network outperforms conditional logistic regression in an age- and sex-matched case control study. BMC Musculoskelet Disord 14:1. https://doi.org/10.1186/1471-2474-14-207
11. Ho-Le TP, Center JR, Eisman JA, et al (2017) Prediction of hip fracture in post-menopausal women using artificial neural network approach. Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2017:4207–4210. https://doi.org/10.1109/EMBC.2017.8037784
12. Ferizi U, Honig S, Chang G (2019) Artificial intelligence, osteoporosis and fragility fractures. Curr Opin Rheumatol 31:368–375. https://doi.org/10.1097/BOR.0000000000000607
13. Hegeman JH, Willemsen G, van Nieuwpoort J, et al (2004) [Effective tracing of osteoporosis at a fracture

and osteoporosis clinic in Groningen; an analysis of the first 100 patients]. Ned Tijdschr Geneeskd 148:2180–2185

14.   DIS open data. https://www.opendisdata.nl/. Accessed 7 Feb 2020

15.   Dziak JJ, Henry KL (2017) Two-Part Predictors in Regression Models. Multivariate Behav Res 52:551–561. https://doi.org/10.1080/00273171.2017.1333404

16.   R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

17.   Cox DR (1972) Regression Models and Life-Tables. J R Stat Soc Ser B 34:187–220

18.   Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Stat 2:841–860. https://doi.org/10.1214/08-AOAS169

19.   Katzman JL, Shaham U, Cloninger A, et al (2018) DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18:24. https://doi.org/10.1186/s12874-018-0482-1

20.   Schoenfeld D (1982) Partial Residuals for The Proportional Hazards Regression Model. Biometrika 69:239. https://doi.org/10.2307/2335876

21.   Steyerberg EW (2019) Clinical Prediction Models. Springer International Publishing, Cham

22.   Shepherd BE, Rebeiro PF (2017) Assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines. J Acquir Immune Defic Syndr 74:e60–e63. https://doi.org/10.1097/QAI.0000000000001221

23.   Breiman L (2001) Random forest. Mach Learn 45:5–32. https://doi.org/10.1017/CBO9781107415324.004

24.   Claesen M, De Moor B (2015) Hyperparameter Search in Machine Learning. 10–14

25.   Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc 1–15

26.   Moradi R, Berangi R, Minaei B (2020) A survey of regularization strategies for deep models. Springer Netherlands

27.   Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830. https://doi.org/10.1145/2786984.2786995

28.   Fotso S (2019) PySurvival: Open source package for Survival Analysis modeling

29.   Kim DW, Lee S, Kwon S, et al (2019) Deep learning-based survival prediction of oral cancer patients. Sci Rep 9:6994. https://doi.org/10.1038/s41598-019-43372-7

30.   Forgetta V, Keller-Baruch J, Forest M, et al (2018) Machine Learning to Predict Osteoporotic Fracture Risk from Genotypes. bioRxiv 413716. https://doi.org/10.1101/413716

31.   Nissinen T (2019) Convolutional neural networks in osteoporotic fracture risk prediction using spine DXA images - Master's thesis Computer Sciences, University of Eastern Finland

32.   Kruse C, Eiken P, Vestergaard P (2017) Machine Learning Principles Can Improve Hip Fracture Prediction. Calcif Tissue Int 100:348–360. https://doi.org/10.1007/s00223-017-0238-7

33.   Ensrud KE, Lui L-Y, Taylor BC, et al (2009) A comparison of prediction models for fractures in older women: is more better? Arch Intern Med 169:2087–94. https://doi.org/10.1001/archinternmed.2009.404

34.   Johansson H, Siggeirsdóttir K, Harvey NC, et al (2017) Imminent risk of fracture after fracture Europe PMC Funders Group. Osteoporos Int 28:775–780. https://doi.org/10.1007/s00198-016-3868-0

35.   Reber KC, König HH, Becker C, et al (2018) Development of a risk assessment tool for osteoporotic fracture prevention: A claims data approach. Bone 110:170–176. https://doi.org/10.1016/j.bone.2018.02.002

36.   Egan M, Jaglal S, Byrne K, et al (2008) Factors associated with a second hip fracture: a systematic review. Clin Rehabil 22:272–82. https://doi.org/10.1177/0269215507081573

37.   de Klerk G (2017) Osteoporosis, identification and treatment in fracture patients

38.   Haentjens P, Autier P, Collins J, et al (2003) Colles fracture, spine fracture, and subsequent risk of hip fracture in men and women. A meta-analysis. J Bone Joint Surg Am 85:1936–43. https://doi.org/10.2106/00004623-200310000-00011

39.   Zhao D, Cheng P, Zhu B (2016) Epilepsy and fracture risk: A meta-analysis. Int J Clin Exp Med 9:564–569

40.   Perna S, Avanzato I, Nichetti M, et al (2017) Association between dietary patterns of meat and fish consumption with bone mineral density or fracture risk: A systematic literature. Nutrients 9:. https://doi.org/10.3390/nu9091029

41.   Wells GA, Cranney A, Peterson J, et al (2008) Alendronate for the primary and secondary prevention of osteoporotic fractures in postmenopausal women. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD001155.pub2