

Machine Learning Principles Can Improve Hip Fracture Prediction

Christian Kruse^{1,2,5}  · Pia Eiken^{3,4} · Peter Vestergaard^{1,2}

Received: 6 September 2016 / Accepted: 5 December 2016 / Published online: 14 February 2017
© Springer Science+Business Media New York 2017

Abstract Apply machine learning principles to predict hip fractures and estimate predictor importance in Dual-energy X-ray absorptiometry (DXA)-scanned men and women. Dual-energy X-ray absorptiometry data from two Danish regions between 1996 and 2006 were combined with national Danish patient data to comprise 4722 women and 717 men with 5 years of follow-up time (original cohort $n = 6606$ men and women). Twenty-four statistical models were built on 75% of data points through k-5, 5-repeat cross-validation, and then validated on the remaining 25% of data points to calculate area under the curve (AUC) and calibrate probability estimates. The best models

were retrained with restricted predictor subsets to estimate the best subsets. For women, bootstrap aggregated flexible discriminant analysis (“bagFDA”) performed best with a test AUC of 0.92 [0.89; 0.94] and well-calibrated probabilities following Naïve Bayes adjustments. A “bagFDA” model limited to 11 predictors (among them bone mineral densities (BMD), biochemical glucose measurements, general practitioner and dentist use) achieved a test AUC of 0.91 [0.88; 0.93]. For men, eXtreme Gradient Boosting (“xgbTree”) performed best with a test AUC of 0.89 [0.82; 0.95], but with poor calibration in higher probabilities. A ten predictor subset (BMD, biochemical cholesterol and liver function tests, penicillin use and osteoarthritis diagnoses) achieved a test AUC of 0.86 [0.78; 0.94] using an “xgbTree” model. Machine learning can improve hip fracture prediction beyond logistic regression using ensemble models. Compiling data from international cohorts of longer follow-up and performing similar machine learning procedures has the potential to further improve discrimination and calibration.

Electronic supplementary material The online version of this article (doi:[10.1007/s00223-017-0238-7](https://doi.org/10.1007/s00223-017-0238-7)) contains supplementary material, which is available to authorized users.

✉ Christian Kruse
ckruse@dcm.aau.dk

Pia Eiken
Pia.Eiken@regionh.dk

Peter Vestergaard
pev@dcm.aau.dk

¹ Department of Endocrinology, Aalborg University Hospital, Moelleparkvej 4, 9000 Aalborg, Denmark

² Department of Clinical Medicine, Aalborg University, Sdr. Skovvej 15, 9000 Aalborg, Denmark

³ Department of Cardiology, Nephrology and Endocrinology, Nordsjaellands Hospital, Dyrehavevej 29, 3400 Hilleroed, Denmark

⁴ Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark

⁵ Department of Endocrinology, Aalborg University Hospital, Hobrovej 19, 9100 Aalborg, Denmark

Keywords Osteoporosis · Fracture · Machine learning · FRAX · Prediction

Introduction

Correctly identifying individuals who will and will not sustain osteoporosis-related fractures is becoming increasingly important with ageing populations and increasing costs of anti-resorptive and anabolic treatment options. Substantial work has been done to estimate fracture risk in different populations (Table 1), but with exceptions [1, 2], linear models such as logistic regression have been predominant. Model performance metrics such as receiver operator

Table 1 Descriptive statistics, men and women with or without 5-year fracture occurrence

	Men, fracture mean \pm SD (<i>n</i>) Ω : median (IQR) (<i>n</i> =47)	Men, no fracture mean \pm SD (<i>n</i>) Ω : Median (IQR) (<i>n</i> =670)	Women, fracture mean \pm SD (<i>n</i>) Ω : median (IQR) (<i>n</i> =293)	Women, no fracture mean \pm SD (<i>n</i>) Ω : median (IQR) (<i>n</i> =4429)
Age, years	69.3 (59.9; 79.6)***	61.8 (50.3; 72.5)	74.5 (65.5; 81.5)***	59.7 (51.1; 70.1)
DXA: Trochanteric BMD (g/cm ²)	0.57 \pm 0.12 $\Omega\Omega\Omega$	0.68 \pm 0.13	0.51 \pm 0.12 $\Omega\Omega\Omega$	0.62 \pm 0.13
DXA: Intertrochanteric BMD (g/cm ²)	0.84 \pm 0.15 $\Omega\Omega\Omega$	1.01 \pm 0.18	0.79 \pm 0.17 $\Omega\Omega\Omega$	0.95 \pm 0.18
DXA: Ward's Triangle BMD (g/cm ²)	0.36 \pm 0.14 $\Omega\Omega\Omega$	0.50 \pm 0.17	0.4 (0.3; 0.4)***	0.5 (0.4; 0.6)
DXA: Femoral Neck BMD (g/cm ²)	0.60 \pm 0.13 $\Omega\Omega\Omega$	0.72 \pm 0.14	0.57 \pm 0.11 $\Omega\Omega\Omega$	0.69 \pm 0.13
DXA: Total Hip BMD (g/ cm ²)	0.73 \pm 0.13 $\Omega\Omega\Omega$	0.87 \pm 0.15	0.68 \pm 0.13 $\Omega\Omega\Omega$	0.81 \pm 0.16
DXA: Hip T-Score, SD	-2.41 \pm -1.01 $\Omega\Omega\Omega$	-1.34 \pm 1.18	-2.56 \pm 1.09 $\Omega\Omega\Omega$	-1.43 \pm 1.26
DXA: Hip Z-Score, SD	-1.58 \pm 1.02 $\Omega\Omega\Omega$	-0.70 \pm 1.15	-0.85 \pm 1.14 $\Omega\Omega\Omega$	-0.39 \pm 1.16
DXA: L1 BMD (g/cm ²)	0.83 \pm 0.19 Ω	0.88 \pm 0.17	0.73 \pm 0.18 $\Omega\Omega\Omega$	0.81 \pm 0.17
DXA: L2 BMD (g/cm ²)	0.87 \pm 0.17 Ω	0.95 \pm 0.18	0.82 \pm 0.19 $\Omega\Omega\Omega$	0.90 \pm 0.19
DXA: L3 BMD (g/cm ²)	0.90 \pm 0.18 Ω	0.96 \pm 0.18	0.87 \pm 0.19 $\Omega\Omega\Omega$	0.94 \pm 0.19
DXA: L4 BMD (g/cm ²)	0.97 \pm 0.23	0.99 \pm 0.20	0.93 \pm 0.21 $\Omega\Omega\Omega$	0.98 \pm 0.19
DXA: Total Spine BMD (g/cm ²)	0.90 \pm 0.18 Ω	0.95 \pm 0.17	0.84 \pm 0.18 $\Omega\Omega\Omega$	0.91 \pm 0.17
DXA: Lumbar Spine T-Score, SD	-1.77 \pm 1.62 Ω	-1.30 \pm 1.55	-1.91 \pm 1.61 $\Omega\Omega\Omega$	-1.29 \pm 1.56
DXA: Lumbar spine Z-score, SD	-0.92 \pm 1.72	-0.67 \pm 1.61	0.10 \pm 1.69	-0.01 \pm 1.57
Dentist Expenses, Weekly DKK (Post)	7.0 (0;19.2)	6.4 (0;14.5)	6.1 (0;17.1)***	8.5 (2.5;17.5)
Medication Expenses, Total DKK (Prior)	9,859 (1,589; 19,197)	7,291 (1,568; 17,832)	9,436 (2,634; 19,971)***	5,271 (1,196; 14,870)
Primary Sector Consulta- tions, Daytime (Post Period)	0.5 (0.3; 0.7)	0.4 (0.2; 0.7)	0.5 (0.3; 0.8)***	0.4 (0.2; 0.7)
Medication expenses, weekly DKK (post)	98.8 (27.4; 157.8)	76.2 (27.3; 189.1)	104.1 (41.6; 187.2)***	62.5 (20.4; 142.4)
GP Expenses, Total DKK (Prior)	13,116 (3228; 124,745)***	100,147 (21,711; 264,920)	15,127 (3817; 183,548)***	132,178 (49,075; 287,360)
Primary Sector Consulta- tions, Unknown (Prior)	8.0 (2.0; 16.0)***	1.0 (0; 8.0)	5.0 (0; 18.0)***	1.0 (0; 7.0)
Dentist expenses, total DKK (prior)	2381 (515; 42,702)	24,640 (0; 83,363)	1363 (0; 41,200)***	43,861 (754; 102,158)

BMD bone mineral density, DKK Danish kroner, DXA dual-energy X-ray absorptiometry, SD standard deviation

*** $p < 0.005$; ** $p < .01$; * $p < .05$

$\Omega\Omega\Omega p < 0.005$; $\Omega\Omega p < .01$; $\Omega p < .05$

characteristics (ROC) are often omitted, instead focusing on individual effect sizes of risk factors. A separation between model metrics obtained from *goodness-of-fit* models and independent validation datasets is also not universally performed.

Supervised machine learning, a “big data” field of applying advanced predictive models and assessing their relative predictive power, has been scarcely applied to osteoporosis and fracture risk research despite the abundance

of retrospective and prospective data available from multiple cohorts. In a transparent way, machine learning can determine the optimal predictive model and its generalizability to the general population, rank the most important predictors and diagnose issues that hinder improved model performance. The same models can be applied to calibrate probabilities attained from these predictions.

We sought to apply such learning principles to an exhaustive dataset of approximately 75,000 predictors

obtained from a combination of regional and national Danish patient data to predict a hip fracture in both men and women. Besides dual-energy X-ray absorptiometry (DXA) bone mineral density (BMD) values commonly used in fracture prediction studies (Table 1), our predictors also include primary health sector use, education level, comorbidity, medication use and biochemistry information. All data were collected and extracted automatically.

Materials and Methods

The Danish National Patient Registry and Related Databases

Information related to health and epidemiology has been collected in Denmark since the founding of the Cancer Registry in 1943 [3]. From 1977 to 1996, respectively, data regarding hospital admissions [4], surgeries and medication use [5] have been collected automatically and provided for academic research and quality assurance. Governance of Danish health data is held by the National Board of Health Data (“Sundhedsdatastyrelsen”) [6].

For medication reimbursements, data consist of purchasing date, package size, Anatomical Therapeutic Chemical (ATC) classification system code and WHO-defined daily dose (DDD). These data are automatically recorded and transmitted by all pharmacies when a prescription is collected. Non-collected prescriptions do not figure in these data. Apart from cosmetics, common weak pain medication and agents against certain addictions (e.g. nicotine), almost all medications require a prescription to be reimbursed in Denmark. Data on hospital admissions consist of in- or outpatient settings, ICD10 code diagnoses, admission and discharge dates. They are collected automatically when a patient enters as an in- or outpatient in relation with the hospital system. Monthly primary sector reimbursement costs and number of visits to different functions [general practitioner (GP), physical therapy, dentistry, etc.] are also recorded. A separate entity, the Danish Civil Registration System [7], has collected national epidemiology and socioeconomic data since 1968 regarding birthdates, sex, dates of death, highest education level, job description and income for the Danish population. Migration information is collected at this entity based on the date of migration, type of migration and origin/destination country.

In this study, we defined the occurrence of a hip fracture from this database by admissions or emergency room visits where the associated ICD10 code defined hip and/or femoral region fractures (M809B, S72, S720, S721, S721A, S721B, S722, S723, S724B, S724C, S727, S728A and S729). The occurrence date was that of admission or the first visit to the emergency room.

Study Design

For this study, inclusion criteria were all men and women who had undergone a DXA scan at the Department of Endocrinology of both Aalborg and Aarhus University Hospitals, Denmark, between 1996 and 2006 with at least 5 years of complete follow-up time from the scan date until 31 December 2011. The included sites cover an administrative region of approximately 1.8 million Danes. The total population was 5.37 million in the autumn of 2011 [8].

The first set of DXA scan data consisting of area, BMC and BMD of the hip (trochanteric, intertrochanteric, Ward’s triangle, FN and total hip) and LS (L1 through L4 and total LS) was selected for each individual and paired with data from the National Board of Health Data from a database of 3 million individuals between 1 January 1996 and 31 December 2011. Start of observation was defined as the date of the first recorded DXA scan, while end of follow-up was defined as either the date of death, emigration, 31 December 2011 or 5 years after the scanning date. A *prior period* was defined as 2 years prior the scan date, while a *post period* was defined from the scan date to fracture or end of follow-up.

Biochemical information consisted of statistically summarized measures such as mean, range, minimum and maximum values for each available component during the *prior* and *post* periods. For a separately defined *closest period* 100 days before the scan date, the value for each component measured closest to the scan date was selected. Medication use was estimated as reimbursed pieces and unstandardized doses for each 5 and 7 digit ATC code during the *prior* and *post* periods. Total medication costs were calculated for the two periods. Categorical presences of all highest hierarchal ICD-10 codes during the *prior* and *post* periods were established categorically, and the maximum length in years from occurrence to the scan date was established for the *prior* period. Charlson comorbidity index [9] total scores were computed and dichotomous presence of individual categories (malignancy, congestive heart failure, etc.) established. Yearly income for the year prior the scan date was established in nominal Danish kroner, not adjusted for inflation or purchasing parity (DKK, 1€ = 7.43 DKK). Primary sector visit count and costs were established for each visit category (e.g. GP, physical therapy, dentist) during both the *prior* and *post* periods. Highest education level at the scan date was established by groups of “never finished primary school”, “finished primary school”, “finished secondary school”, “vocational training”, “vocational bachelor” (“Professionsbachelor”), “academic bachelor”, “university graduate” or “post-graduate”. Job description was grouped by categorical independent variables. Ethnicity was determined as either Danish or by the first origin country of the migration data. *Post* period predictors were relativized to

terms per week of observation. Age, sex, height and BMI were collected from the DXA scan visits.

The machine learning procedures were performed for the resulting predictors and associated with the categorical hip fracture or non-fracture outcome.

Machine Learning Introduction

Machine learning refers to the overlapping discipline of computer science and computational statistics. Aside from genetics [10, 11], this field is only sporadically used in medicine, but is becoming more relevant as increasing computational power and larger datasets enable the use of predictive models with greater accuracy than linear models. Within machine learning, supervised learning refers to the prediction of a pre-defined outcome using different models on the same dataset. The reason for the improved accuracy of the models lies in the so-called *complexity* or *tuning* parameters that result in varying model simplicity and complexity. Machine learning principles optimize these tuning parameters to perform well on not only the model dataset, but also on independent data points.

An intuitive comparison can be done with logistic regression, which is traditionally modelled as decision boundaries from a linear model (first-degree polynomial). From a given dataset of predictors, only one model will be fitted in a linear context. The decision boundary could also be modelled from a second, third or n -th-degree polynomial, or from a constant 0-degree polynomial. A polynomial of a higher degree will model the data very precisely compared to a lower-degree polynomial. However, when applied to new and independent data points, a model with a higher degree polynomial can show very poor performance compared to the original precision metrics of the *goodness-of-fit* model, while the model with a lower-degree polynomial can perform poorly, but comparable to the *goodness-of-fit* model. These observations are caused by *model variance* and *model bias*. A too complex model will be poorly generalizable to new data points by *inappropriately high model variance* (“overfit”), whereas a too simple model will have both poor *goodness-of-fit* and generalized precision through *inappropriately high model bias* (“underfit”). By creating multiple models across several complexity parameters and testing them through cross-validation, the optimal degree of model complexity that generalizes well can be computed.

The machine learning principles generally involve splitting a dataset into *training* (75% of data points) and *test/validation* (25% of data points) subsets, where the *training* datasets are used to build and validate models of different complexities several times on inner splits or *folds* (“cross-validation”) [12–14]. Models are thereby “trained” to the best balance between predictive capability (*training dataset error*) and lowest risk of poor

generalization (*test dataset error*) by summarizing performance measures from *cross-validations* [15, 16]. Receiver operator characteristics [17, 18] or Cohen’s kappa k are used to evaluate classification models (categorical outcome), while R^2 or root mean square error (RMSE) can be used for continuous outcomes [[19–21] he probabilities estimated from the classification models can be calibrated by *fitting* them to the outcome and re-predicting the probabilities with the adjusting models. By comparing the prevalence of occurrences within bins of probabilities (e.g. “Did 80% of the individuals with probabilities of 80% sustain the outcome?”), the soundness of the probabilities can be visualized and described through different metrics, e.g. the Lemeshow–Hosmer *goodness-of-fit* [[22]

Several model categories and individual models exist, but for classification problems such as this study, they are generally grouped as *linear/discriminant models*, *non-parametric models* and *tree-based models*. The former group includes traditional logistic regression, linear discriminant analysis and partial least squares. Non-parametric models include powerful but abstract models such as *neural networks* and *support vector machines*, where the former involves a hidden layer of features computed from the input features, and the latter involves decision boundaries that have the largest margin of separation between outcomes. Tree-based models are based on classification and regression trees that traverse predictors for cut-off values to separate two outcomes best by entropy. Recursively, predictors are again traversed through the resulting groups until a tree-like structure of paths and nodes is constructed and can no longer be expanded. Solitary trees are often regarded as weak learners, but by procedures such as ensembling (random forest, bootstrap aggregation) or boosting (subsequently remodelling with different weights to hard-to-predict data points), the performance of the models can be markedly improved by lowering bias and to some degree variance as well. Strengths and limitations differ between the models, i.e. the inclusion of missing data, collinearity issues and the need for centring and scaling data to find the global minimum of error of the cost functions.

The different models differ greatly from interpretable but generally poorly performing to powerful but abstract. The general approach is to find the predictive ceiling of the dataset and then ascertain if more interpretable models are non-significantly worse than this, allowing for easier interpretation and utility in clinical use or further causal studies. Ranked variable importance can be extracted to list which predictors were most important for differentiating outcome from non-outcome, and by backwards feature elimination or forward “greedy” addition methods, the optimal subset of predictors can be estimated.

Machine Learning Procedure

The machine learning procedure in this study first involved the removal of *near-zero variance* predictors (frequency-ratio 19, unique cut-off 10%) and predictors with more than 60% missing observations, resulting in a “non-complete case” dataset including missing values. Outliers were not removed due to their relevance in specific models. By random forest imputation (1500 trees through 5 iterations), a separate “imputation” dataset was created. The two datasets were split into “training” datasets of 75% of cases and “test” datasets of the remaining 25% of cases. For models that feature implicit feature selection and accept missing values, the “non-complete case” dataset was used, while models not accepting missing values were subjected to the “imputation” dataset.

Twenty-four computed models were trained through k-5, 5-repeat cross-validation to maximize ROC. The models, their applicable tuning parameters and short descriptions are provided in Supplementary Material 1. The final models were then applied to the independent test dataset to attain probabilities for each outcome and compared to the observed test dataset outcomes; a ROC curve was computed to establish the combination of discriminatory power (AUC) with a 95% bootstrapped confidence interval. Sensitivity and specificity were calculated at both the Youden *J*-index [23] and a 50% probability cut-off. Calibration analyses were performed by relativizing the observed prevalence within bins of predicted probabilities, e.g. “Did 80–90% of the individuals with predicted probabilities between 80–90% sustain the outcome?”. The uncalibrated probabilities were calibrated through Platt scaling [24], isotonic regression [25], neural networks, a Naïve Bayes approach and support vector machines (radial, linear and polynomial kernels). The optimal calibration approach for each model was chosen by visualization without Lemeshow–Hosmer metrics. The best models for the two examined groups were selected subjectively by a combination of discriminatory power (AUC) and calibration of probabilities.

Variable importance [26] was extracted from the best performing model, and sequentially by a “greedy forward selection method”, models were built on the training dataset with variables ranked 1 through *n* (e.g. variables 1–2, variables 1–3, variables 1–4, ..., variables 1–30) on the same model type, then applied for prediction on the test dataset. ROC curves between predicted and observed were estimated, followed by AUC calculation with 95% CI by bootstrapping ($n = 20,000$ samples).

Statistics

For descriptive statistics, mean and standard deviation (SD) was calculated for parametric data, and median and 25th–75th percentile for non-parametric data. Distributions were evaluated visually by QQ-plots, histograms and kernel smoothing. Summary statistics on biochemical samples were computed as the number of measurements, and mean value, SD, minimum, maximum, range, lower and upper 95% confidence interval (CI) true mean boundaries were calculated for the “prior” and “post” periods by component and individual. Three statistical significance levels were defined as $p < 0.05$, $p < 0.01$ and $p < 0.001$.

Statistical analyses were performed using SAS (Version 9.4, 64-bit, SAS Institute Inc., Cary, NC) and R (The R Foundation, version 3.2.3, Vienna, Austria). R packages are listed in the *Machine Learning Procedure* section.

Results

A total of 4722 women and 717 men with 5 years of follow-up time met inclusion criteria from a pool of 6606 men and women scanned between 1996 and 2006. Median ages of men and women with and without fracture occurrence were 69.3 and 61.8 years and 74.5 and 59.7, respectively, with strong statistical tendencies for lower hip BMD values, greater medication expenses and lower general practitioner expenses in the fracture groups (Table 1). Within the 5-year observation period, 293 women (6.62%) and 47 men (6.55%) sustained a hip fracture. A total of 74,989 predictors were present in the original dataset, reduced to 1255 after removal of *near-zero variance* predictors and predictors with inappropriate levels of missing data. After trimming of the original predictor set, records were 100% complete for hip DXA data, between 96% (L4) and 100% (L1) complete for LS DXA data, between 41.2% (plasma levels of immature granulocytes) and 82% complete (plasma levels of total calcium) for biochemistry data and 100% complete for comorbidity, medication use and socioeconomic data.

In the female cohort, the “bootstrap aggregated flexible discriminant analysis” model (“bagFDA”) showed the best balance between discriminatory power and calibrated probabilities. The greatest numerical AUC was achieved by the “eXtreme Gradient Boosting” model with an AUC of 0.92 [0.89; 0.94], but due to poor probability calibration, the “bagFDA” model was chosen. This model achieved a test AUC of 0.91 [0.88; 0.94] and at the Youden probability cut-off, sensitivity and specificity were 88 and 81%, respectively. Uncalibrated probabilities overestimated occurrences throughout all bins but could be calibrated very well with a Naïve

Bayes approach, except for underestimating probabilities in the 90–100% bins (Supplementary Material 2). Calibrated AUC using the Naïve Bayes adjustment was unchanged at 0.91. The most important predictors for this model were different types of primary sector use, markers of diabetes (Fig. 1) mellitus, trochanteric and intertrochanteric BMD measurements (Table 2). The best predictor subset consisted of 11 predictors that achieved a test AUC of 0.906 [0.88; 0.93] (Fig. 2a). These predictors were (1) GP Expenses, Weekly DKK (Prior Period), (2) 1+Episode of Increased P-Glucose (mmol/L) (True/False), (3) 1+Samples of P-Glucose (mmol/L) (Post period), (4) Dentist Expenses, weekly DKK (Post period), (5) Fusidic acid (S01AA13), Weekly DDD Reimbursed (Prior period), (6) 1+Samples of P-Thyroid Stimulating Hormone (Post Period), (7) Dentist Expenses, Total DKK (Prior Period), (8) 1+Dentist Consultation (Prior Period), (9) DXA: Intertrochanteric BMD (g/cm^2), (10) Acute GP Service consultations, No of (Post Period) and (11) DXA: Trochanteric BMC (g).

In the male cohort, the “eXtreme Gradient Boosting” model achieved the best balance between discriminatory power and calibration. The model achieved a test AUC of 0.89 [0.82; 0.95] with a sensitivity of 100% and specificity of 69% at the Youden probability cut-off. The model’s uncalibrated probabilities above 70% performed very poorly with great overestimations, and although they improved with a Naïve Bayes approach, the probabilities still underestimated occurrences noticeably for probabilities above 75% (Supplementary Material 3). Calibrated AUC was lowered to 0.77. The most important predictors in this cohort were biochemical markers of cholesterol metabolism and liver function (alanine aminotransferase and albumin), primary sector use (general practitioners and dentists), oral penicillin reimbursement, DXA-based hip Z-score and diagnosis codes associated with osteoarthritis (M17.1 - Unilateral primary osteoarthritis of knee) and age. A model consisting of the top 9 predictors achieved a test AUC of 0.86 [0.78; 0.94]. These predictors were (1) 1+Episode of Hypocholesterolemia (True/False, Post Period); (2) GP Consultations, Number of Unknown Character (Type 0, Prior Period); (3) Phenoxymethylpenicillin (J01CE02), Pieces Reimbursed (Prior Period); (4) Dentist Expenses, Total DKK (Prior Period); (5) Existing Diagnosis M17.1 (Unilateral primary osteoarthritis of knee), Time Since First (Years or No); (6) DXA: Total Hip Z-Score (SD); (7) Existing Diagnosis S80.0 (Contusion of Knee), Time Since First (Years); (8) GP Expenses, Weekly DKK (Post Period); (9) P-LDL Cholesterol (mmol/L), Range (Period Period) (Table 3).

Discussion

We present the first combined use of several advanced predictive models from the field of supervised machine learning on hip fracture prediction in a population of DXA-scanned men and women. We document that ensemble tree-based models that use *boosting* and *bootstrap aggregation* approaches can improve discriminatory capabilities on independent subjects and provide acceptable calibrated probabilities with the best reliability for the female cohort. We believe these performance metrics can be further improved through compilations of existing international datasets and longer observation periods.

In the two examined cohorts, predictive performance and calibrated probabilities were good for both men and women, but best in the larger female group with an AUC value of 0.91. Generally, existing fracture prediction studies (Table 4) use the well-established Fracture Risk Assessment Tool (FRAX[®]) model to estimate 10-year calibrated probabilities and compare them to observed outcomes. The outcomes in the FRAX[®] model are hip fracture and major osteoporotic fractures modelled to a combination of DXA measurements and risk factors. The pioneering work on the FRAX[®] model by Kanis et al. [27, 28] echoes our approach of building models and validating them on independent datasets, and in those studies of cohorts from different geographical regions and age groups, validation AUC metrics of 0.66 were reached for models built on risk factors only and 0.74 for models built on risk factors and BMD (age and sex-standardized Z-score). While comparing our 5-year follow-up to 10-year estimates is troublesome, a numerical AUC of 0.91 for females is an improvement of the high FRAX[®] validation AUC achieved by Azagra et al. in the Spanish FRIDEX cohort (AUC 0.88 [0.82; 0.95]) [29] and in similar Danish results by Friis-Holmberg et al [30]. (AUC of 0.86 [0.81; 0.92]) using phalangeal BMD. This also indicates in a very strong way that the discriminatory ceiling for hip fractures is even higher than 0.91 with the predictors used in our study, as a 10-year observation period will inevitably include more fractures and balance the outcomes towards higher potential sensitivity.

Our study illustrates the potential improvements to osteoporosis research that can be achieved with machine learning, and also to discuss the computational burdens and pitfalls that this discipline presents in comparison with traditional logistic and Cox regression. The intuition of model complexities for the same predictor subset is challenging when we as clinicians are accustomed to generalized linear models that model outcomes to predictors in exactly one way. The pitfalls of not validating predictive models are not universally known either, as we disappointingly note how some hip fracture prediction studies (Table 1) solely reported *goodness-of-fit* performance measures without

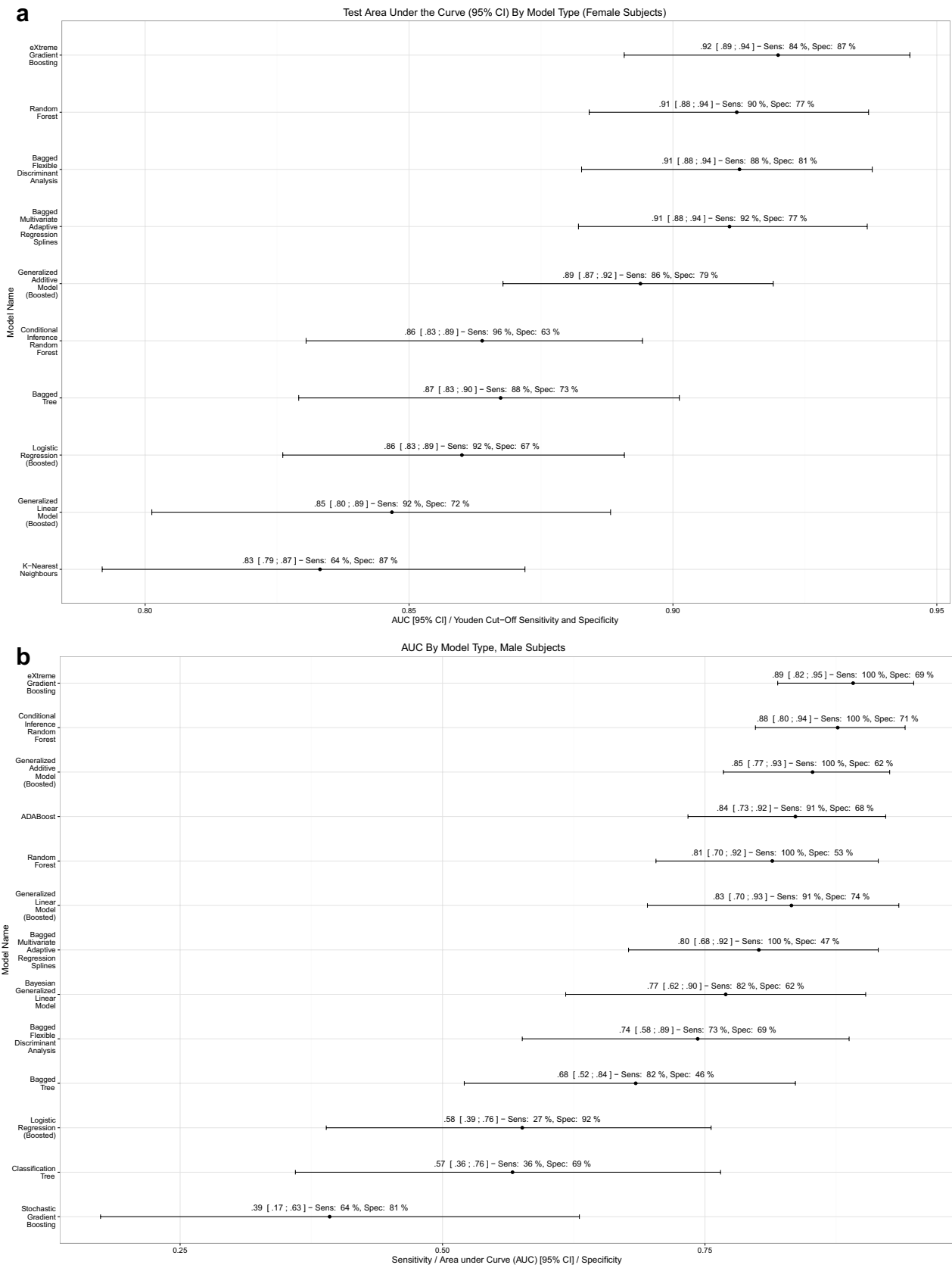


Fig. 1 **a** Test dataset model performances specified by AUC (area under the curve), sensitivity and specificity female subjects. **b** Test dataset model performances specified by AUC (area under the curve), sensitivity and specificity male subjects

Table 2 Female subjects

Predictor	Relative importance (standardized 0–100)	Training AUC (95% CI)	Test AUC (95% CI)
1 GP Expenses, Weekly DKK (Prior Period)	100		
2 1+ Episode of Increased P-Glucose (mmol/L) (True/False)	80	0.769 [0.73; 0.81]	0.777 [0.72; 0.83]
3 1+ Samples of P-Glucose (mmol/L) (Post Period)	69	0.825 [0.81; 0.84]	0.822 [0.76; 0.88]
4 Dentist Expenses, Weekly DKK (Post Period)	60	0.850 [0.83; 0.87]	0.851 [0.80; 0.90]
5 Fusidic acid (S01AA13), Weekly DDD Reimbursed (Prior Period)	53	0.869 [0.85; 0.88]	0.874 [0.83; 0.91]
6 1+ Samples of P-TSH (Post Period)	47	0.870 [0.86; 0.88]	0.872 [0.83; 0.91]
7 Dentist Expenses, Total DKK (Prior Period)	40	0.858 [0.84; 0.87]	0.884 [0.85; 0.92]
8 1+ Dentist Consultation (Prior Period)	35	0.875 [0.86; 0.89]	0.893 [0.86; 0.93]
9 DXA: Intertrochanteric BMD (g/cm ²)	26	0.882 [0.87; 0.89]	0.902 [0.87; 0.93]
10 Acute GP Service Consultations, No of (Post Period)	14	0.887 [0.87; 0.90]	0.903 [0.87; 0.93]
11 DXA: Trochanteric BMC (g)	2	0.886 [0.87; 0.90]	0.906 [0.88; 0.93]
12 DXA: Trochanteric BMD (g/cm ²)	0.2	0.886 [0.87; 0.90]	0.905 [0.88; 0.93]

Relative predictor importance and progressive feature addition modelling. Stochastic gradient boosting models fit with 1 through n (e.g. predictors 1–2, 1–3, 1–4, 1: n), then applied to independent test dataset with Youden J probability cut-off. Bootstrapped AUC w/ 95% CI and bootstrapped ROC test progressively through the models

AUC area under the curve, *Bold* final model of variables 1 through 12 marked in with resulting training and test AUC, *Relative predictor importance* calculated from entropy measures, standardized from 0 to 100

BMC bone mineral content, BMD bone mineral density, DDD WHO-defined daily doses, DKK Danish kroner, DXA dual-energy X-ray absorptiometry, GP general practitioner, TSH thyroid stimulating hormone, WHO World Health Organization

external validation. These two cornerstones of machine learning indicate several likely obstacles to establishing machine learning as a practice in osteoporosis. When a predictive model is not validated on independent data points, the performance estimates from the training models are likely too optimistic. We illustrate this in our progressive feature addition runs (Fig. 2a, b) where some loss of AUC is experienced when a model is generalized or tested on independent datasets. The underlying concept is the balance between model *bias* (inappropriately low complexity) and model *variance* (inappropriately high complexity) which is a cornerstone of predictive modelling. When simply adapting a pre-selected subset of predictors with no internal or external validation and experiencing low error metrics on training datasets, there is a great risk of high model variance that leads to poor performance on new data points. Through the k-5, 5-repeat internal cross-validation procedure we perform with each model type and relevant tuning parameters, we seek to limit these risks of inappropriate variance and bias [31, 32] by selecting the optimal model complexity through intense computations. With 100 tuning parameters, this necessitates 2500 individual models for one model type alone, before the optimal model complexity is chosen. Yet this also provides several positive aspects, as it allows us to transparently document how well our predictive capability is, as well as limiting the bias associated with human interference in predictor selection. For several models, the machine learning procedure also allows for automatic feature selection by discarding

irrelevant predictors, which allowed us to reduce 75,000 predictors to 11 for women and 9 for men.

Beyond the improvements in prediction attainable by the systematic learning approach, we also designed this study to illustrate how this technology can fit into current electronic health records (EHR) databases and carry a great potential for personalized medicine. Data underlying every predictor in this study were either collected automatically (i.e. gender and age from the social security number, medication use by prescription reimbursements) or as feature engineered predictors of automatically collected data (i.e. mean plasma low-density lipoprotein). This eliminates the need for questionnaires from the patients and the associated technicalities (e.g. what constitutes ‘rheumatoid arthritis’ when the answer must be yes or no?). Through a combination of databases and statistical software, the predictors and models of this study can be recreated in other EHR systems. Ho-Le et al. [33] recently exemplified how machine learning and the random forest model could be used on single-nucleotide polymorphism data to rank individual genetic factors in fracture risk. As in our study, this illustrates the added value data can attain through machine learning in several fields. The further potential of reapplying data is to create a system that continuously remodels hip fracture risk to account for temporal changes in risk factors and to provide personalized risk alterations through simulations. We deliberately included the *post period* in our modelling, as this will allow us to combine existing knowledge

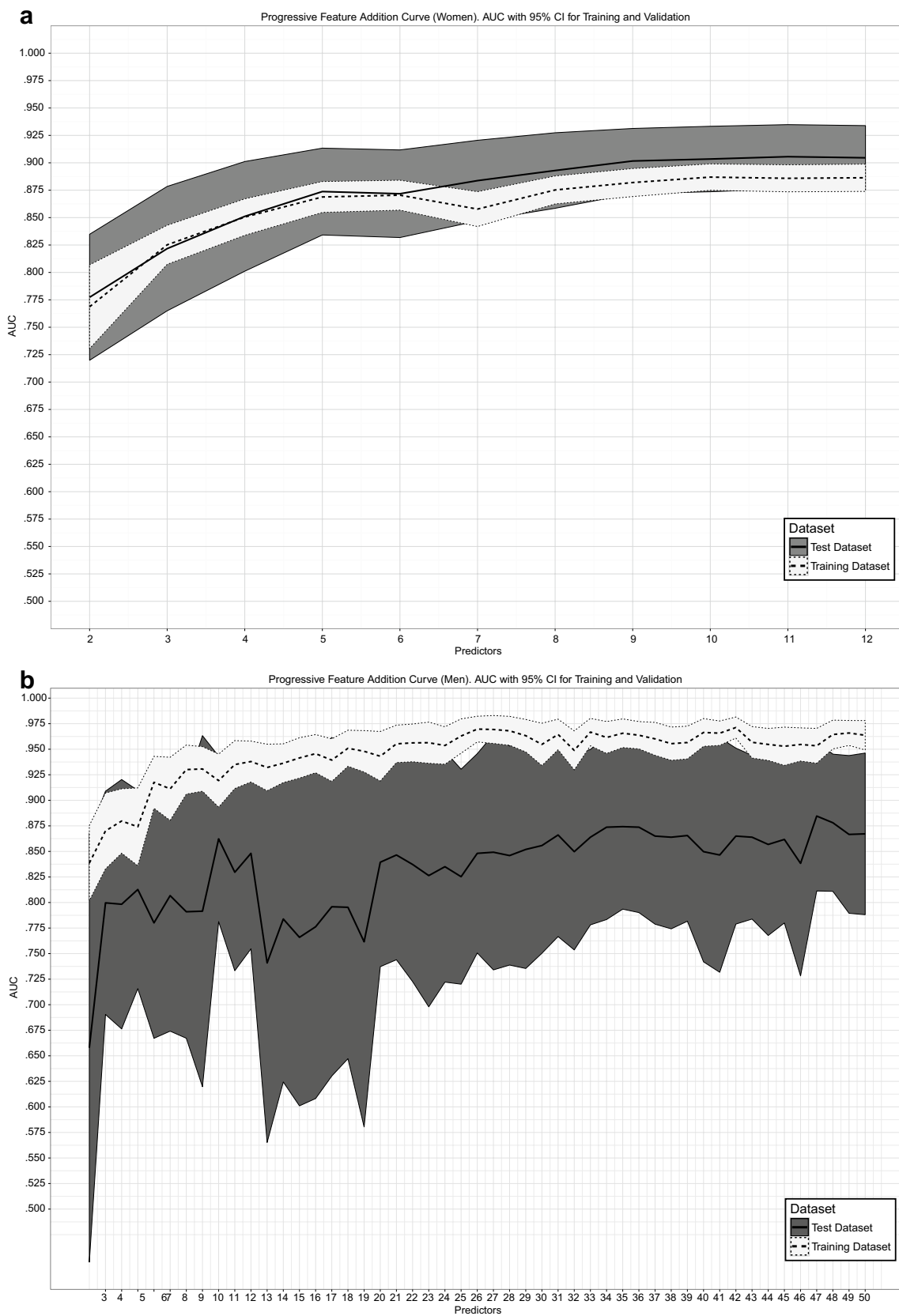


Fig. 2 a Training and test dataset performances specified by AUC (area under the curve) with 95% CI for predictor subsets 1 through n (e.g. 1:1, 1:2, 1:3 ... 1:30) female subjects. **b** Training and test data-

set performances specified by AUC (area under the curve) with 95% CI for predictor subsets 1 through n (e.g. 1:1, 1:2, 1:3 ... 1:30) male subjects

Table 3 Male subjects

Predictor	Relative importance (standardized 0–100)	Training AUC (95% CI)	Test AUC (95% CI)
1 1+Episode of Hypocholesterolemia (True/False, Post Period)	100	0.839 [0.80; 0.88]	0.658 [0.45; 0.87]
2 GP Consultations, No. of Unknown Character (Type 0, Prior Period)	52	0.870 [0.83; 0.91]	0.800 [0.69; 0.91]
3 Phenoxyethylpenicillin (J01CE02), Pieces Reimbursed (Prior Period)	48	0.880 [0.85; 0.91]	0.798 [0.68; 0.92]
4 Dentist Expenses, Total DKK (Prior Period)	44	0.874 [0.84; 0.91]	0.813 [0.72; 0.91]
5 Existing Diagnosis M17.1 (Unilateral primary osteoarthritis of knee), Time Since First (Years or No)	41	0.918 [0.89; 0.94]	0.780 [0.67; 0.89]
6 DXA: Total Hip Z-Score (SD)	37	0.911 [0.88; 0.94]	0.807 [0.67; 0.94]
7 Existing Diagnosis S80.0 (Contusion of Knee), Time Since First (Years)	31	0.930 [0.91; 0.95]	0.791 [0.67; 0.91]
8 GP Expenses, Weekly DKK (Post Period)	31	0.931 [0.91; 0.95]	0.792 [0.62; 0.96]
9 P-LDL Cholesterol (mmol/L), Range (Period Period)	29	0.919 [0.89; 0.95]	0.862 [0.78; 0.94]
10 Existing Diagnosis M17.1 (Unilateral primary osteoarthritis of knee) (True/False)	29	0.935 [0.91; 0.96]	0.830 [0.73; 0.93]
11 P-Lymphocytes ($10^9/L$), Minimum Value (Post Period)	27	0.938 [0.92; 0.96]	0.848 [0.75; 0.94]
12 DXA: Total Hip BMD (g/cm^2)	26	0.932 [0.91; 0.95]	0.741 [0.57; 0.92]
13 Samples Drawn of P-Eosinophiles ($10^9/L$), No of	24	0.936 [0.92; 0.96]	0.784 [0.62; 0.94]
14 P-ALAT (U/L), Chronically Increased (True/False, Post Period)	22	0.941 [0.92; 0.96]	0.766 [0.60; 0.93]
15 Phenoxyethylpenicillin (J01CE02), Weekly DDD Reimbursed (Post Period)	20	0.946 [0.93; 0.96]	0.776 [0.61; 0.94]
16 P-Albumin (g/L), Mean Value (Prior Period)	19	0.939 [0.92; 0.96]	0.796 [0.63; 0.96]
17 Dentist Expenses, Weekly DKK (Post Period)	18	0.951 [0.93; 0.97]	0.795 [0.65; 0.94]
18 Age at DXA Scan (years)	18	0.948 [0.93; 0.97]	0.762 [0.58; 0.94]
19 Existing Diagnosis S61.0 (Open wound of thumb without damage to nail) (True/False)	16	0.943 [0.92; 0.97]	0.839 [0.74; 0.94]

Relative predictor importance and progressive feature addition modelling. eXtreme Gradient Boosting models fit with 1 through n (e.g. predictors 1–2, 1–3, 1–4, 1: n), then applied to independent test dataset with uncalibrated probabilities. Bootstrapped AUC w/ 95% CI and bootstrapped ROC test progressively through the models

AUC area under the curve, **Bold** final model of variables 1 through 20 marked in with resulting training and test AUC, *Relative Predictor Importance* calculated from entropy measures, standardized from 0 to 100

ALAT alanine aminotransferase, DDD WHO-defined daily doses, DKK Danish kroner, DXA dual-energy X-ray absorptiometry, GP general practitioner, LDL low-density lipoprotein, U international units, WHO World Health Organization

(e.g. existing GP expense data) with simulations of *post period* options (e.g. *higher* or *lower post period* GP expenses) to assess how fracture probabilities can change with different behaviours and treatment choices. As an example, this technology will be able to assess the probability of fractures when prescribing bisphosphonates, denosumab, teriparatide or other osteoporosis treatments to individuals of different comorbidity profiles, concurrent medication use and primary healthcare sector use. It is our strong opinion that the field of osteoporosis and fracture risk assessment should involve statistical learning to a greater degree, and we suggest that the existing datasets be compiled and a systematic machine learning approach of cross-validated training, probability calibration and external validation be performed using the statistical models of this study.

Strengths and Limitations

This study was designed and implemented to limit the human involvement and risk of information bias from the databases as much as possible. The aim was to separate the data from the interpretation. The original dataset was exhaustive of all available diagnoses, medication types, occupation types and biochemical information that could be found in the ICD-10 and ATC systems. For the national data, we believe our implementation could be seen as if data were collected prospectively from 1 January 1996 onwards. As an example, we present “pieces reimbursed”, but cannot state if these pieces were indeed consumed, at what frequencies and which doses. The definition of hip fractures was done by ICD-10 codes which have previously been shown to have high accuracy rates [34]. The

Table 4 Studies of hip fracture prediction in women (men and women where applicable)

Authors and cohort	Country	Patient group	Exposure	Outcome	Goodness-of-fit or validation	AUC (95% CI)
Kruse et al. (Aalborg & Aarhus Cohort, 2016)	Denmark	4722 women	DXA measurements w/ national patient data	Hip fracture	Validation	0.92 [0.89; 0.94]
		717 men	DXA measurements w/ national patient data	Hip fracture	Validation	0.89 [0.82; 0.95]
Kälvesten et al. (SOF, 2016)	USA	5278 women, 65+ years	Age w/ FRAX ®	Hip fracture	Goodness-of-fit	0.70 [0.67; 0.73]
			Age w/ FN BMD & FRAX ®	Hip fracture	Goodness-of-fit	0.76 [0.73; 0.78]
Caillet et al. (EPIDOS, 2015)	France	7598 women, 75+ yrs	FN BMD w/ Risk Factors	Hip fracture	Goodness-of-fit	0.72 [0.70; 0.75]
Friis-Holmberg et al. (Danish Health Examination Survey, 2014)	Denmark	7552 women, 40–90 years	FRAX ® w/o Phalangeal T-score	Hip fracture	Validation	0.860 [816; 0.903]
			Phalangeal T-score	Hip fracture	Validation	0.834 [0.777; 0.890]
			FRAX ® w/ Phalangeal T-score	Hip fracture	Validation	0.862 [0.809; 0.916]
Praemaor et al. (SOF, 2013)	Multi-site	1,509 women, mean age 72.3 yrs	FRAX ®	Hip fracture	Validation	0.69 [0.67; 0.71]
			FRAX ® w/ FN BMD	Hip fracture	Validation	0.73 [0.71; 0.76]
Azagra et al. (FRIDEX, 2012)	Spain	770 women, 40–90 yrs	FRAX ® w/o FN T-score	Hip fracture	Validation	0.888 [0.824; 0.952]
			FRAX ® w/ FN T-score			0.849 [0.737; 0.962]
Sambrook et al. (GLOW, 2011)	Multi-site	19,586 women, postmenopausal	FRAX ®	Hip fracture	Validation	0.78
Hundrup et al. (Danish Nurse, 2010)	Denmark	15,648 women, postmenopausal	Risk factors (Robbins et al. algorithm)	Hip fracture	Validation	0.82
Robbins et al. (WHI, 2007)	USA	93,676 women, 50–79 years	Risk factors	Hip fracture	Validation	0.80 [0.77; 0.82]
Kanis et al. (2007)	Multi-site	46,340 men and women	Risk factors w/o FN Z-score	Hip fracture	Validation	0.57 (PERF). 77 (Miyama)
			Risk factors w/ FN Z-score	Hip fracture	Validation	0.65 (Geelong I). 81 (Geelong 2)
McGrother et al. (Melton Cohort, 2002)	UK	1289 women (70+ years)	Risk factors	Hip fracture	Cross-validation	0.73
Burger et al. (Rotterdam Study, 1999)	Netherlands	5208 women and men	Risk factors	Hip fracture	Goodness-of-fit	0.83
			Risk factors w/ FN BMD	Hip fracture	Goodness-of-fit	0.88

AUC area under the curve, EPIDOS Multi-site study, France 1992–1996, FRIDEX Fracture Risk factors and bone Densitometry type central dual X-ray Cohort, Barcelona Spain, GLOW Global Longitudinal Study Of Osteoporosis in Women, multi-site study in ten countries, Melton Cohort primary sector cohort in Melton, Leicestershire. SOF study of Osteoporosis Cohort, multi-site, WHI Women's Health Initiative, multi-site American cohort, W/ with, W/O without, UK United Kingdom, USA United States of America

models were trained by tuning parameters appropriate to the dimensionality of predictors and data points and only expanded if model performance had not reached a ceiling.

The main weakness in our study involves the interpretability of our findings and the selected population. As we describe a population of DXA-scanned women, we do not include subjects who were not DXA scanned and therefore

at a potentially higher risk than scanned individuals. The inclusion criteria of 5 years of follow-up time resulted in exclusion of several scanned individuals from the original data, but as this is not related to the outcome or the exposure, we do not expect particular bias from this limitation. Our aim was to describe the relevance of DXA scan measurements and the individual regions of interest. As an

example, the use of primary sector consultations is likely a composite measure of health psychology and behaviour that we cannot describe further. The biochemical data should be regarded as retrospective and there are substantial limitations to this portion of the data, as several sample types were likely drawn by indication rather than scientific exploration. An effort to circumvent this by applying a factor of “sample drawn” versus “sample not-drawn” limited this problem. Traditional risk factors such as smoking and alcohol intake could not be included in our model. Finally, computational requirements limited our study, as the dimensionality of 4400 data points with originally 74,989 predictors provided memory limitations. The computations were performed in a data server setting of 32 CPU cores with 512 GB of RAM, but this still required us to remove near-zero variance predictors and predictors with an inappropriate level of missing data. This approach is controversial as it can remove predictors that are uncommon but important [35], but is currently a necessity unless expensive data centre solutions are provided for public sector use.

Conclusion

We conclude that hip fracture risk can be modelled with high discriminative performance for men (Test AUC of 0.89 [0.82; 0.95], sensitivity 100%, specificity 69% at the Youden probability cut-off) and particularly for women (Test AUC 0.91 [0.88; 0.94], sensitivity 88%, specificity 81% at the Youden probability cut-off) using advanced predictive models. Ensemble models using *bootstrap aggregation* and *boosting* performed best in both cohorts, and probabilities can generally be calibrated well with a Naïve Bayes approach, although poor for high probability estimates in men. Models of 11 predictors for women and 9 for men with combinations of DXA BMD measurements and primary sector use achieved the highest numerical AUC values. Further improvements in predictive capability are likely possible with compilations of more data points and longer observation periods. We strongly suggest the use of machine learning principles to model hip fracture risk, and we welcome an effort to compile existing datasets and perform advanced predictive modelling.

Acknowledgements We acknowledge Statistics Denmark for providing data and a server platform for data analysis. The Obel Family Foundation of Aalborg, Denmark, and the Department of Clinical Medicine at Aalborg University, Denmark, are acknowledged for funding the PhD fellowship of Dr. Christian Kruse. Grant Numbers are not applicable in Denmark.

Author Contributors Christian Kruse designed the study and performed data management, modelling, model validation, statistical analysis, graphical presentations and manuscript preparation of first draft of the paper. He is guarantor. Pia Eiken and Peter Vestergaard

performed revisions and final approval of the manuscript draft. All authors revised the paper critically for intellectual content and approved the final version. All authors agree to be accountable for the work and to ensure that any questions relating to the accuracy and integrity of the paper are investigated and properly resolved.

Compliance with Ethical Standards

Conflict of interest CK has received travel grants from Eli Lilly, Otsuka Pharmaceutical and is a speaker for Novartis and Otsuka Pharmaceutical. PE is an advisory board member with Amgen, MSD and Eli Lilly and at the speakers bureau with Amgen and Eli Lilly, stocks from Novo Nordisk A/S. PV has received unrestricted grants from MSD and Servier, and travel grants from Amgen, Eli Lilly, Novartis, Sanofi-Aventis and Servier.

Human and Animal Rights and Informed Consent The procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008.

References

1. Lin CC, Ou YK, Chen SH, Liu YC, Lin J (2010) Comparison of artificial neural network and logistic regression models for predicting mortality in elderly patients with hip fracture. *Injury* 41(8):869–873
2. Jin H, Lu Y, Harris ST et al (2004) Classification algorithms for hip fracture prediction based on recursive partitioning methods. *Med Decis Mak* 24(4):386–398
3. Sundhedsdatastyrelsen, Cancerregistret. <http://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygedomme-laegemidler-og-behandlinger/cancerregisteret>
4. Sundhedsdatastyrelsen, Landspatientregistret. <http://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygedomme-laegemidler-og-behandlinger/landspatientregisteret>
5. Sundhedsdatastyrelsen, Lægemedelstatistikregisteret. <http://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/sygedomme-laegemidler-og-behandlinger/laegemiddelstatistikregisteret>
6. Sundhedsdatastyrelsen. <http://sundhedsdatastyrelsen.dk/da>
7. CPR-registret. <http://www.cpr.dk>
8. Statistics Denmark. <http://www.dst.dk/da/Statistik/emner/befolkning-og-befolkningsfremskrivning/folketal.aspx>
9. Sundararajan V, Henderson T, Perry C, Muggivan A, Quan H, Ghali WA (2004) New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J Clin Epidemiol* 57(12):1288–1294
10. Mitra AK, Mukherjee UK, Harding T et al (2016) Single-cell analysis of targeted transcriptome predicts drug sensitivity of single cells within human myeloma tumors. *Leukemia* 30(5):1094–1102
11. Sharma GB, Robertson DD, Laney DA, Gambello MJ, Terk M (2016) Machine learning based analytics of micro-MRI trabecular bone microarchitecture and texture in type 1 Gaucher disease. *J Biomech* 49(9):1961–1968
12. Cohen G, Hilario M, Pellegrini C, Geissbuhler A (2005) SVM modeling via a hybrid genetic strategy. A health care application. *Stud Health Technol Inform* 116:193–198

13. Kim JH (2009) Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. *Comput Stat Data Anal* 53(11):3735–3745
14. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Jt Conf Artif Intell* 14:1137–1145
15. Simon R, Radmacher MD, Dobbin K, Mcshane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95(1):14–18
16. Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics* 21(15):3301–3307
17. Altman DG, Bland JM (1994) Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 309(6948):188
18. Brown C, Davis H (2006) Receiver operating characteristics curves and related decision measures: a tutorial. *Chemom Intell Lab Syst* 80(1):24–38
19. Kvalseth T (1985) Cautionary note about R2. *Am Stat* 39(4):279–285
20. Hawkins DM, Basak SC, Mills D (2003) Assessing model fit by cross-validation. *J Chem Inform Comput Sci* 43(2):579–586
21. Martin J, Hirschberg D (1996) Small sample statistics for classification error rates I: error rate measurements. Department of Informatics and Computer Science Technical Report
22. Lemeshow S, Hosmer DW (1982) A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 115(1):92–106
23. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3(1):32–35
24. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif* 10(3):61–74
25. Zadrozny, C Elkan (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD
26. Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77(4):802–813
27. Kanis JA, Oden A, Johnell O et al (2007) The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int* 18(8):1033–1046
28. Kanis JA, Johnell O, Oden A, Dawson A, De laet C, Jonsen B (2001) Ten year probabilities of osteoporotic fractures according to BMD and diagnostic thresholds. *Osteoporos Int* 12(12):989–995
29. Azagra R, Roca G, Encabo G et al (2012) FRAX® tool, the WHO algorithm to predict osteoporotic fractures: the first analysis of its discriminative and predictive ability in the Spanish FRI-DEX cohort. *BMC Musculoskelet Disord* 13:204
30. Friis-holmberg T, Rubin KH, Brixen K, Tolstrup JS, Bech M (2014) Fracture risk prediction using phalangeal bone mineral density or FRAX(®)?-a Danish cohort study on men and women. *J Clin Densitom* 17(1):7–15
31. Hawkins DM (2004) The problem of overfitting. *J Chem Inform Comput Sci* 44(1):1–12
32. Van Der Putten P, Van Someren M (2004) A bias-variance analysis of a real world learning problem: the CoIL challenge 2000. *Mach Learn* 7(1–2):177–195
33. Ho-le TP, Center JR, Eisman JA, Nguyen HT, Nguyen TV (2016) Prediction of bone mineral density and fragility fracture by genetic profiling. *J Bone Miner Res*. Doi: [10.1002/jbmr.2998](https://doi.org/10.1002/jbmr.2998)
34. Vestergaard P, Mosekilde L (2002) Fracture risk in patients with celiac Disease, Crohn's disease, and ulcerative colitis: a nationwide follow-up study of 16,416 patients in Denmark. *Am J Epidemiol* 156(1):1–10
35. Zorn C (2005) A solution to separation in binary response models. *Political Anal* 13(2):157–170