Prediction of a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis: A predictive model as a risk assessment tool for future fractures

Health Technology and Services Research (HTSR)

Master's Thesis Health Sciences

ZGT Almelo, April 2020

B.C.S. de Vries

1225146

University of Twente

Health Sciences

SUPERVISORS

University of Twente:

Dr. C.G.M. Groothuis-Oudshoorn

Dr. C. Seifert

Ziekenhuisgroep Twente (ZGT):

Dr. J.H. Hegeman

Ing. J. Geerdink

Drs. W.S. Nijmeijer

“Artificial intelligence will not replace physicians. However, physicians who use Artificial Intelligence will replace those who do not.”

Bertalan Meskó MD, Director of the Medical Futurist Institute (2018)

“The greatest opportunity offered by AI is not reducing errors or workloads, or even curing cancer: it is the opportunity to restore the precious and time-honored connection and trust—the human touch—between patients and doctors.”

Eric Topol MD, in his book Deep Medicine (2019)

Preface

You are currently reading the master thesis I conducted to graduate for the master Health Sciences at the University of Twente, titled: "*Prediction of a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis: A predictive model as a risk assessment tool for future fractures*". I aimed to fill in a clinical need at the department of traumatology at Ziekenhuisgroep Twente (location Almelo), by developing a fracture prediction tool in the context of a Fracture and Osteoporosis Outpatient Clinic. Curious if I succeeded? The answer is in this report.

For me, this thesis is successful regardless of the answer on the above stated question. Beforehand, I wanted to get an understanding of predictive modelling and improve my very basic programming skills to some more useful skills. I now feel more confident on these topics and think I have gained important insights which I can use as a future physician. Besides, I once again experienced that I have a very active learning style. However, I also discovered that this is not always the best approach in research and some reflection beforehand may come in handy.

I would like to thank my supervisors for their guidance and support in this project. Karin, thank you for giving me a solid start by sharing your knowledge on both model development and imputation methods, and for always being open to my questions. Han, thank you for the opportunity to perform this research in the ZGT and your guidance on the clinical relevance of the model. Christin, thank you for enlightening me in the world of Machine Learning and your help regarding Python. Jeroen, thank you for your help on the data-extraction in Hix and the fun we had in our meetings. Wieke, thank you for your suggestions, I appreciate that you despite the distance to Groningen provided useful feedback.

Lastly, I would like to thank my family and friends. A huge thank you to my parents who, not only have me on their payroll for almost more than 9 years, but also unconditionally support me. Sophie, thank you for putting my feet back on the ground regarding both my aims of this thesis and myself in general. You are the best and loveliest sparring partner anyone could wish for.

Bram

Drachten, April 2020

Abstract

Purpose:

Major osteoporotic fractures (MOFs), defined as fractures from hip, wrist, spine and humerus, can have serious consequences regarding morbidity and mortality. Artificial intelligence gives new opportunities for fracture prediction and may aid in targeting preventive interventions to patients at risk of MOF. Primary objective of this study is to develop and compare several models, capable of predicting risk of MOF as a function of time in patients who already sustained a fracture.

Methods:

Patients aged >50 visiting the osteoporosis screening clinic after sustaining a fracture were included in this retrospective pilot study. We compared discriminative ability (concordance-index) for time to MOF prediction of a Cox regression model, a Random Survival Forest (RSF) model and an artificial neural network (ANN)-DeepSurv model. Missing data was imputed using multiple imputation by chained equations (MICE) and RSF's imputation function. Analysis were performed for the total cohort and a subset; 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), no significant difference was determined between the models. In the osteopenia subset, Cox regression outperformed RSF (p=0.026 and p=0.036) and ANN-DeepSurv (p=0.042) with a c-index of 0.625 (0.562-0.689). Cox regression was used to develop a MOF risk calculator in this subset.

Conclusion:

We present adequate discriminative performance of different prediction models and show Cox regression to outperform RSF and ANN in osteopenia patients. We developed a user-friendly tool for risk calculation of subsequent MOF in patients with osteopenia.

Table of contents

List of tables

List of figures

List of abbreviations

Adam: Adaptive moment estimation, - 20 -

AI: Artificial intelligence, - 3 -

ANN: Artificial neural network, - 3 -

AUC: Area under the receiver operating characteristic curve, - 13 -

BMD: Bone mineral density, - 1 -

C-index: Concordance statistic, - 13 -

COCP: Combined oral contraceptive pill, - 17 -

DBC: Diagnosis treatment combinations, - 16 -

DXA scan: Dual-energy X-ray absorptiometry, - 2 -

eGFR: Estimated glomerular filtration rate, - 17 -

ESR: Erythrocyte sedimentation rate, - 17 -

GP: General practitioner, - 33 -

IQR: Interquartile range, - 21 -

LASSO: Least absolute shrinkage and selection operator, - 9 -

MAR: Missing at random, - 4 -

MCAR: Missing completely at random, - 4 -

MICE: Multivariate imputation by chained equations, - 5 -

MNAR: Missing not at random, - 4 -

MOF: Major osteoporotic fracture, - 1 -

RSF: Random survival forest, - 3 -

SD: Standard deviation, - 1 -

SQL: Structured Query Language, - 16 -

TSH: Thyroid stimulating hormone, - 17 -

1 Introduction

Osteoporosis is a major cause of bone fractures in elderly. Globally, 158 million individuals older than 50 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 osteoporotic fracture is expected to have doubled in 2045 due to progressive ageing [1]. Major osteoporotic fractures (MOFs), defined as fractures from 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 mortality and morbidity [4]. Twenty to thirty-five percent of patients admitted to the hospital with a fractured neck of the femur die within one year [5]. Besides, studies suggest that costs of osteoporosis compared to other diseases are relatively high [6,7]. An international report quantified the cost of osteoporosis and its consequences as 98 billion euro for the European Union in 2010, almost as high as the burdensome disease dementia (105 billion). The costs in the Netherlands alone for osteoporosis were quantified at 2.7 billion and are expected to increase to 3.5 billion in 2025 [6].

The WHO recognized and defined osteoporosis in 1991 as: “A disease characterised by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and a consequent increase in fracture risk” [8]. This definition was extended in 1994 with the establishment of four general diagnostic categories, based on bone mineral density (BMD), and still used nowadays:

1. Normal bone mass, a BMD value of 1 standard deviation (SD) from the young adult reference mean.
2. Osteopenia, a BMD value of more than 1 SD below the young adult reference mean but less than 2.5 SD below this value.
3. Osteoporosis, a BMD value of 2.5 SD or more below the young adult reference mean.
4. Severe osteoporosis, a BMD value of 2.5 SD or more below the young adult reference mean in the presence of one or more fragility fractures. [9]

The value of BMD compared to the young adult reference mean is commonly known as the T-score.

Osteoporosis gained attention near the turn of the century and is now widely studied. The University Medical Centre Groningen (UMCG) opened the first Fracture and Osteoporosis outpatient clinic in the Netherlands in 2003 to assess the presence of osteoporosis in patients at risk.

Schurink and colleagues (2007) demonstrated effectiveness of this outpatient clinic presenting a rise in BMD measurements of patients at risk from 14% to 75%. Moreover, they determined that 69% of the patients not screened for osteoporosis actually had an indication for treatment with bisphosphonates according to the osteoporosis guidelines [10]. Furthermore, patients with a prior fracture of any nature are well 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 no fracture, while the relative risk for a subsequent fracture after hip fracture, vertebral fracture or radius/ulna fracture is 3.2 (2.3 – 4.5), 3.0 (2.0 – 4.3), and 2.7 (1.9 – 3.7) compared to no fracture respectively [11]. The current Dutch guideline 'Osteoporosis and fracture prevention' recommends BMD screening by means of a dual-energy X-ray absorptiometry (DXA scan) in patients older than 50 with a recent non-vertebral fracture [12]. As a result, most hospitals nowadays have Fracture and Osteoporosis outpatient clinics.

Treatment decision for prevention of osteoporosis is currently based on general classification of patients. For example, patients with a T-score < -2.5 and/or a vertebral fracture are treated with bisphosphonates, while other patients not meeting this criterion are not. Recently, healthcare made a shift towards targeted prevention using more personalised medicine [13]. Personalised medicine is defined by the Horizon 2020 Advisory Group of the European Commission as: "A medical model using characterisation of individuals' phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention" [14]. In the field of fracture prevention, risk assessment tools aim to personalize future fracture risk to support treatment decisions. 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 [15]. FRAX® uses twelve input parameters and is based on linear and non-linear combinations of risk-factors for future fracture [15]. However, it does not consider several important proven risk factors such as relevant comorbidities, patient's history and prior fall(s), amongst others. 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 [12]. Consequently, the current guideline in the Netherlands does not advice to use the FRAX®-tool or any other fracture risk assessment tool for clinical decision making, but only in risk-communication to patients with osteopenia [12].

Relatively new modelling techniques give rise to new opportunities for fracture prediction as they can handle large numbers of input variables simultaneously. Artificial neural networks (ANNs) are one of these new approaches, showing 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 [16]. Besides, Ho-le et al. (2017) showed that their ANN outperformed k-nearest-neighbour, support vector machine and logistic regression in the prediction of hip fracture in post-menopausal women [17]. Artificial intelligence (AI) based models will potentially capture underlying trends and patterns, making predictions more accurate and therefore more useful in clinical practice [18].

The primary objective of this pilot study is to develop and compare several traditional and non-traditional models, capable of predicting time to event of a subsequent MOF in patients who already sustained a (minor) fracture. The best performing model could possibly serve in a risk-assessment tool in clinical practice for patients visiting a Fracture and Osteoporosis Outpatient Clinic. Such risk-assessment tool may be helpful in targeting preventive interventions for patients at high risk of MOF fracture who do not currently meet the criteria for bisphosphonate treatment. Secondary aim of this study is to identify predictors of subsequent fractures in this population. Cox proportional hazard regression is traditionally used to predict time to event in survival data, while Random Survival Forest (RSF) is a popular machine learning method used for this purpose [19]. Used models in this thesis are therefore Cox proportional hazard regression, RSF and an ANN.

2 Theoretical background

Developing and validating a model requires completion of several steps, nicely outlined by Steyerberg (2019) [20]. Primarily, one must deal with missing values and correctly code the predictors. Next, the model needs to be specified with an appropriate selection of the main effects and an assessment of the assumptions considered. The model is then estimated and the performance of the model determined. Final steps include model validation (internal or external) and presentation of the model to the audience. The latter is out of the scope of this research. Below, the theoretical background of the used methods is outlined. At first, methods to handle missing data are described as the dataset used was incomplete. Secondly, an introduction to survival data is given and three different models able to handle survival data are discussed. Lastly, background regarding evaluation methods for performance of the models are given.

2.1 Dealing with missing data: Multiple imputation

Incomplete data in retrospective studies are frequently occurring problems. It is caused by different factors, such as part of questionnaires not being answered, errors in data registration and outliers. Deleting missing data is a standard approach in handling this problem, known as listwise deletion [21]. However, as different variables might be missing in non-overlapping data, this can result in significant reduction of statistical power. According to Rubin (1976), missing data can be classified into three different categories:

- Missing Completely At Random (MCAR): the probability that data is missing is considered the same for all individual patients.
- Missing At Random (MAR): the probability of missing is the same for all individual patients within groups defined by the observed data.
- Missing Not At Random (MNAR): the probability that data is missing is not considered the same for all individual patients due to reasons that are unknown to the researchers [22].

Different methods can be used to impute missing data, including some simple methods such as mean imputation. These, however, have a serious drawback regarding the produced standard errors in the final analysis after imputation. Multiple imputation, first developed in 1987 by Rubin, is a technique able to solve this problem [21,23].

Van Buuren (2012) described that multiple imputation is generally accepted as the best method to deal with incomplete data and is often being used as benchmark to which newer methods are compared [21]. This simulation-based statistical technique consists of three important steps:

1. Imputation, in which several imputed datasets are created. In every dataset, all missing values are identified and replaced by plausible data values drawn from a modelled distribution. This modelled distribution is constructed using a prediction method combined with noise and parameter uncertainty. This step results in *m* complete datasets, with *m* reflecting the chosen number of datasets. These datasets are identical on the initially available data but differ in the values imputed.
2. Completed-data analysis, where desired models are constructed and analysed for every imputed dataset separately.
3. Pooling, in which results of separate completed-data analysis for every dataset is combined into an overall result. This step applies the so-called Rubin's rules, assuming normally distributed parameters [21].

Above steps are illustrated in Figure 1.

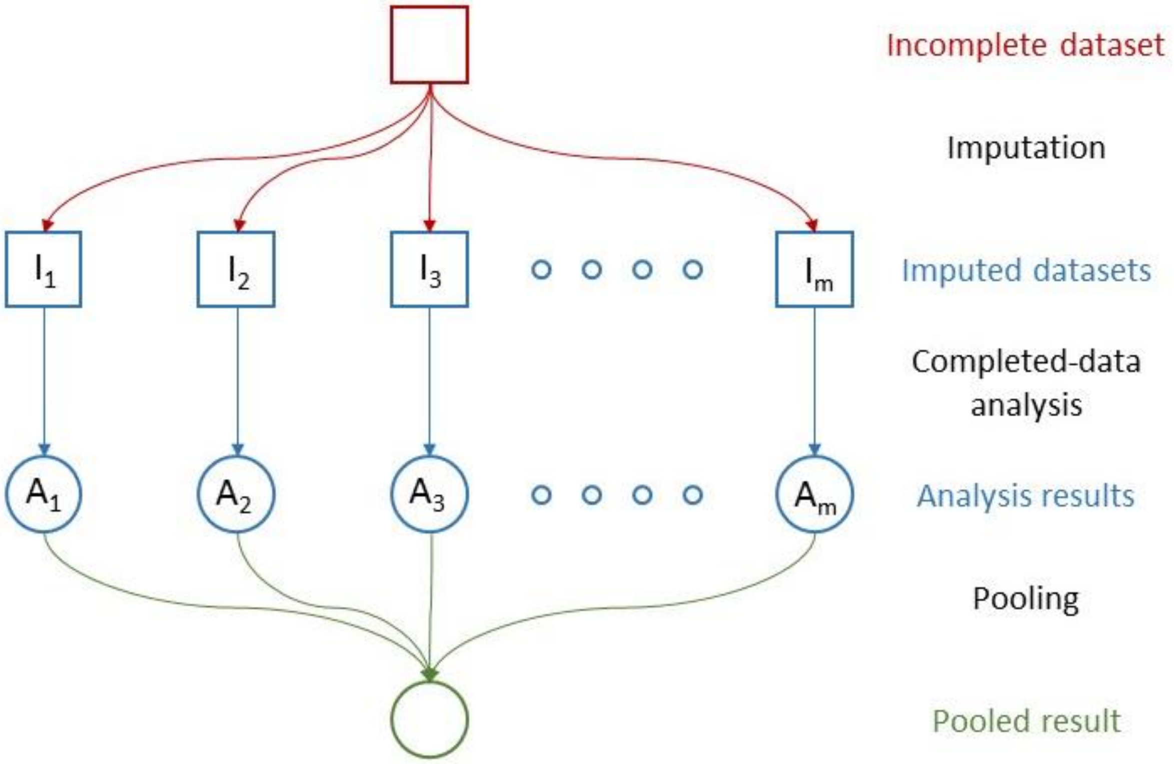


Figure 1: Overview of the steps of multiple imputation. Multiple imputed datasets (I1, I2, I3, …, Im) are created by replacing the missing values in the incomplete dataset with values drawn from a modelled distribution. Secondly, the desired analysis is performed for every completed dataset. Finally, the results of the separate complete-data analysis (A1, A2, A3, …, Am) are pooled into an overall result. The chosen number of datasets is reflected by m.

We used a fully conditional specification approach, also known as multivariate imputation by chained equations (MICE), designed by Van Buuren and Groothuis-Oudshoorn (2000) [24,25]. This method defines a multivariate model through a set of univariate models with specific conditions and imputes incomplete data points by values iterated from this conditional model. It can handle data straightforward when MCAR or MAR is assumed [25]. In this research, missing data as assumed to be MAR.

2.2 Survival analysis

Survival analysis is a statistical technique capable of modelling time to event based on historical data. Survival data consist of two outcomes measures, event (yes/no) and the time till event. This type of data often contains censored data points, meaning that the time-to-event of these data points is unknown. Including these data without altering the statistical model results in an underestimation of the time to event [26]. Censoring can occur in different ways:

1. Right censoring, for individuals in which an event is assumed to occur but occurs beyond the follow-up period. Possible causes of the end of follow-up period are termination of the study period, loss to follow-up or death.
2. Left censoring, for individuals in which the time of the event is not exactly known, but only by its upper limit. The most important cause is a follow-up period with intervals.
3. Interval censoring, for individuals in which the time of the event is not exactly known but its lower and upper limit are.

Right censored data is most common and is properly dealt with by most statistical survival techniques [27].

Survival data can be described by different probabilities. The survival probability *S(t)* is defined as the probability that an individual survives from start of inclusion to a specified future time *t*. The hazard *h(t)* is the probability that an individual, who is under observation at time *t*, has an event at that time. Survival reflects the cumulative non-occurrence of an event, while the hazard is related to the incidence event rate. Both probabilities are related to each other. This relation is described by the following formula:

(1) $h(t) = -\frac{d}{dt}[\log S(t)]$

As there is no simple way to estimate *h(t)*, the cumulative hazard *H(t)* is normally used in statistics. This is the area under the hazard function between times 0 and t, and is related to *S(t)* in the following way:

(2) $H(t) = -[\log S(t)]$

Different methods have been proposed to estimate the cumulative hazard function and more are currently being developed [26]. The three methods used in this thesis are described in the upcoming sections.

2.3 Cox proportional hazard regression

The most well-known technique is Cox proportional hazard regression model, developed in 1972 [28]. This semi-parametric model, which unlike a full-parametric model leaves the dependence on time unspecified defines the hazard function (3) of the Cox model as follows:

(3) $h(t,x,\beta) = h_0(t)e^{x_1\beta_1+x_2\beta_2+\ldots+x_p\beta_p}$

with h_0 defined as baseline hazard function; t as survival time; x_1, x_2, ..., x_p as covariates and β_1, β_2, ..., β_p, as coefficients of the corresponding covariates.

The Cox proportional hazard regression model maximizes the partial likelihood to get the estimate of the coefficient of a covariate. It assumes proportionality of hazards, meaning that the ratio of hazards remains constant over time [29]. The proportionality assumption can be tested using Schoenfeld's residual correlation test [30]. Another important assumption is the linear effect of continuous variables on the log hazard function. This assumption can be tested in several ways, including fractional polynomials or splines and a graphical check using the log hazard ratio [29,31].

2.3.1 Restricted cubic splines

As discussed earlier, Cox proportional hazard regression model assumes a linear effect of continuous variables on the log hazard function. In other words, it is assumed that the effect is consistent for all values of the variable. However, not all continuous covariates might be linearly related to the log hazard function and therefore may lead to an inaccurate model. To use non-linear covariates in a regression model, covariates needs to be transformed. One very flexible form of these transformation are spline functions [20].

Spline functions are polynomials on intervals used for fitting a curve. The higher the order of a spline, the more flexible it is. There are several forms of splines, one of interest for this research is the restricted cubic splines, sometimes referred to in literature as natural splines. The restricted cubic spline is a piecewise polynomial with a high polynomial order and is therefore able to fit functions which are sharply shaped and not correctly fitted by traditional transformations. Knots are used to bend the spline function around. The number of knots is equal to the number of degrees of freedom for restricted cubic splines. The term 'restricted' relates to the tails of the spline, which is constrained to be linear [20,32]. Restricted cubic splines can therefore be used to model non-linear relationships in the Cox model.

2.3.2 Variable selection using LASSO

One of the major risks in modelling is overfitting; when a model describes data of the sample used in the analysis very closely, but does not give valid predictions for new subjects [20]. Important topics related to overfitting are bias and variance. Definitions for bias and variance are clearly described by Ghojogh and Crowley (2019); bias describes how much the mean of the estimate deviates from the original mean, while variance is defined as the average deviation from the mean of the estimate [33]. Four illustrative examples of high and low bias and variance using a standard dartboard example are shown in Figure 2.

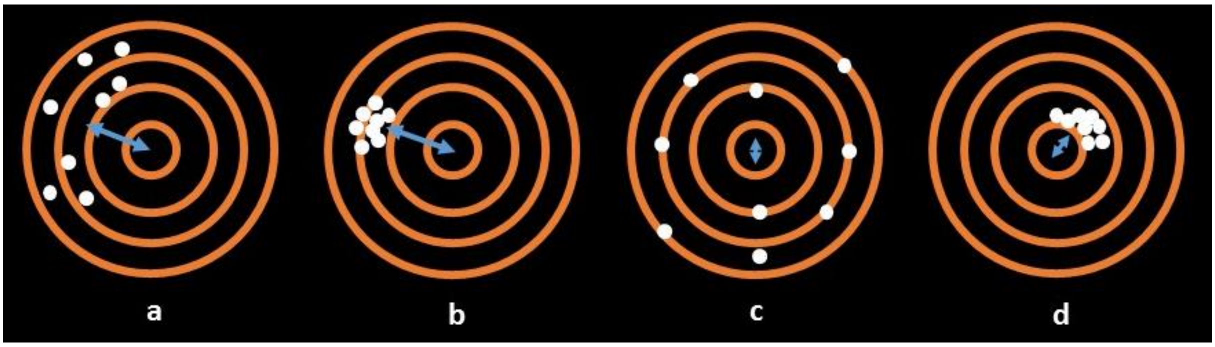


Figure 2: Illustrative examples of bias and variance with: a) high bias and variance, b) high bias and low variance, c) low bias and high variance, d) low bias and low variance.

In overfitting, variance is high while bias is low. Two important causes of overfitting in regression modelling are parameter uncertainty and model uncertainty. Parameter uncertainty results in overestimation of regression coefficients at the extremes of a linear predictor, this phenomenon can be explained by Stein's paradox [20]. Stein (1956) determined that biased estimates are preferable over unbiased estimates to make better predictions in multivariate models [34]. Including some bias in the model, and thus reducing variance, might therefore result in a gain in predictive precision for new subjects outside training data. This issue is commonly referred to as the bias-variance trade-off in literature and is visually illustrated by Dankers et al. (2019) in the book Fundamentals of Clinical Data Science [35]. A copy of this illustration is shown in Figure 3.

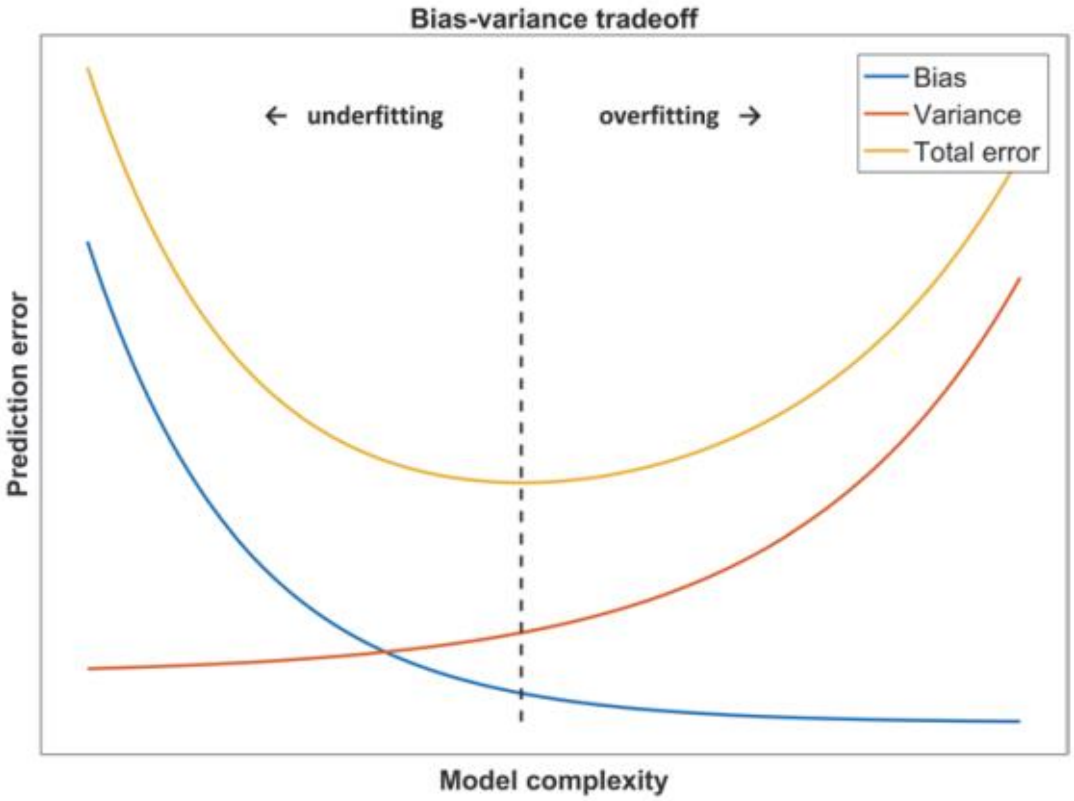


Figure 3: The bias variance trade-off as illustrated by Dankers et al. (2019) [35]. A higher variance results in a more accurate match of the underlying relation in the training set, but can also raise the prediction error. The prediction error is the sum of bias and variance and needs to be minimized.

The other cause of overfitting is model uncertainty, resulting from testimation bias. Testimation bias is overestimation of predictors' effects, resulting from selection methods which only include predictors with a relatively large effect, such as stepwise selection methods. This results in a model with a higher variance and therefore causes overfitting. Steyerberg (2019) advices to limit the use of traditional stepwise selection methods and include improved methods such as the least absolute shrinkage and selection operator (LASSO) [20].

LASSO is a modern estimation technique, which adds bias in the regression coefficients and thereby reduces variance of the model. LASSO does this by penalizing the sum of the absolute values of regression coefficients. This results in some coefficients becoming 0 and therefore being excluded from the model. The challenge however is to find the optimal bias-variance trade-off, resulting in the best predictive ability of the model for new subjects [20]. An optimum for the penalty factor (λ) can be determined by defining the minimum of the partial likelihood deviance using cross validation. λ is usually chosen as the minimum penalty factor plus one standard error as proposed by the author [36].

2.4 Random survival forest

RSF is, compared to the Cox model, a new statistical technique introduced by Ishwaran et al (2008). It is a decision tree-based method capable of estimating the cumulative hazard function of survival data [19]. Unique selling points of decision trees are its explainability and the easy way they can handle qualitative predictors. However, trees themselves are non-robust, meaning a small change in data results in a large change in the final estimate. Therefore they generally have a lower predictive ability than other models [37].

Random forest is a technique combining bootstrap aggregating and decorrelation to build multiple decision trees, thereby reducing variance. Moreover, it can handle both classification and regression problems. Bootstrap aggregating, or in short bagging, is a technique in which multiple bootstrap training data sets are created by taking repeated samples from the training dataset. The model is then trained on these newly created training sets and averaged to define the final prediction. Decorrelation is used to prevent one very strong predictor to annex all these trees, therefore reducing variance. For every split used to construct the decision trees, only a random sample of m predictors is used as split candidates instead of the full set of p predictors. m is mostly chosen as $m = \sqrt{p}$ [37].

RSF, developed by Ishwaran et al. (2008), extends the method of random forest to right-censored survival data. This model constructs a cumulative hazard function for every tree and averages these to obtain an ensembled cumulative hazard function. Besides, it can calculate the variable importance, giving insight in prediction model and its variates. The variable importance of a specific variable is determined by analysing out-of-bag samples. A new model is developed which considers a random split instead of a split based on that specific variable. It therefore excludes this specific variable from the new model [19]. The prediction error of the new model is then compared to the original model. The variable importance is defined as the increase in prediction error of the new model compared to the original model [38]. Lastly, RSFs can impute missing data using a methodology specifically designed for Random Forests by Ishwaran et al (2008) [19].

2.5 Artificial neural network

ANNs originate from mathematical theories used to describe the information processing of a neural system in an animal brain. Already in 1943, McCulloch and Pitts published a theory for so called nerve nets, a forerunner of the artificial neural networks used nowadays [39]. The fast growth in computational capability in the last decades led to new opportunities on this topic and a huge increase in interest of different industries, including healthcare. ANNs are valued for their practical and flexible approach and are particularly useful when data entail complex interactions, violate specific assumptions or contain a large unexplained variance [40].

ANNs use multiple layers to describe the association between input and outcome. Typically, ANNs consist of 3 types of layers:

- Input layer, which receives information from an external source such as a database.
- Hidden layer, which is responsible for the internal processing of the data by the mean of weights. The optimal weights are defined by minimizing the average error between the real outcome and the prediction.
- Output layer, which produces the final output of the model [41].

The type used in this research is a feed-forward neural network, meaning information flow in the model is unidirectional. Both input and output layer always comprise a single layer. When a hidden layer is present, it can consist of single or multiple layers.

Katzman et al (2018) designed a Cox proportional hazards deep neural network, called DeepSurv [42]. It is a feed-forward neural network which estimates the hazard function by analysing the effect of the patient's covariates parameterized by the weights of the network and is therefore able to deal with survival data. The input to the network is the patient's data. The hidden layers are fully connected layers with a non-linear activation function [42]. Dropout is a regularization method used in the hidden layer of the neural network during the training phase. It temporarily removes nodes and their connections at random, thereby reducing overfitting [43] The output layer of DeepSurv comprises an individual node estimating the log-risk function in a Cox proportional hazard model using linear activation [42]. An example of such a network is shown in Figure 4.

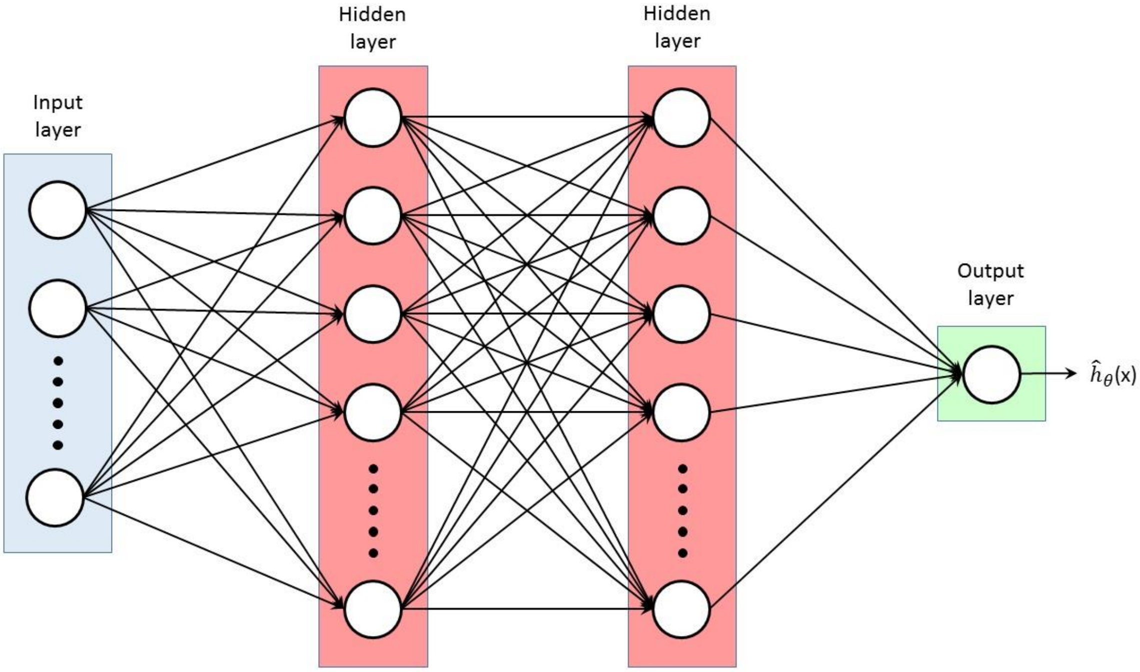


Figure 4: Example of a feed-forward neural network with 2 hidden layers. The output layer estimates the log-risk function in a Cox proportional hazard model.

Gradient descent optimization is used to train the network minimizing the average negative log partial likelihood. The average negative log partial likelihood is a measure of the goodness of fit of the model for given data. The loss function of DeepSurv, the function used to evaluate the set of weights of the network, is defined as:

$$(4)\ \imath(\theta) \coloneqq -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}_\theta(x_i) - \log \sum_{j\in\theta\Re(\mathrm{T}_\imath)} + \lambda \cdot \|\theta\|_2^2$$

Hyper-parameters of the network are depth and size of the network, learning rate, ℓ_2 regularization coefficient, dropout rate, exponential learning rate decay constant and momentum [42].

The choice of hyper-parameters is crucial for the model’s performance, as a good set of hyper-parameters maximizes the discriminative ability of the learning approach. Therefore, hyper-parameter search is an important topic in machine learning research [44]. There are several optimization techniques, with grid search and manual search being most widely used. In grid search, the model is run for all combinations of a manually pre-specified subset of hyper-parameters, returning an evaluation of the model performance. The hyper-parameters from the best performing model are then used for the final model. A serious drawback of grid search is its large computational time, as the model needs to be constructed for all combinations of hyper-parameters [45].

2.6 Evaluation of model performance

The most used statistic for discriminative performance in survival analysis is the concordance statistic (c-index). Concordance is derived from the Wilcoxon-Mann-Whitney U two sample rank test and is defined as the probability that the prediction goes in the same direction as the actual data. In case of a binary outcome measure, e.g. in logistic regression, the area under the (receiver operating characteristic) curve (AUC) equals the c-index [31,37,38].

As discussed earlier, in survival analysis one needs to deal with both time-to-event data and censored data. Every subject at time t experiencing an event is compared to all other comparable subjects at risk. It uses three different sets of paired comparisons:

- Event vs. non-event, comparing the predicted probability of the event and the non-event.
- Event vs. event, comparing the predicted probability of both events with respect to the duration of time till the event. The patient who developed an event earlier is supposed to have a higher predicted probability for an event.
- Event vs. censored, comparing the predicted probability of the event and the censored case only when the censored time is longer than the time till event.

Next, these components are combined into an overall measure of discrimination. The c-index is the proportion of pairs in which the subject experiencing the event has a higher calculated risk of the event compared to the other subject [48]. Like AUC, a c-index of 0.5 is comparable to random guessing while a c-index of 1.0 reflects perfect discrimination.

2.6.1 Validation

Besides a good performance of the model on training data, a high predictive ability for new patients is essential. Model validation is a measure to assess model performance for new subjects and can be done both internally and externally. Internal validation assesses validity of the model on data in the same setting as the model was trained on, while external validation uses samples which are fully independent from training data. As external validation is out of scope of this pilot study, we will focus on internal validation using cross-validation.

Cross-validation is a widely known validation technique, splitting data in a training- and a test-set and repeating this several times. It thereby makes sure there is no overlap in the test sets. The model is developed using training data, and consecutively evaluated using test data. The size of the training- and test-sets depend on the amount of repetitions used. For example, in a 10-fold cross validation, the training-set is split in 90% and 10% for the training- and test-set respectively. This process is repeated ten times making sure all patients have served once as a test object for the model [20]. An important risk in cross-validation is data leakage, defined as using data or information during model generation from the test set which leads to overfitting. For example, when doing both hyper-parameter search and model evaluation using cross-validation, the test sets of the hyper-parameter search and final model evaluation may overlap partly [35]. A solution to this violation of assumption is nested cross-validation, which uses an inner loop to find the optimal parameters of the model and an outer loop for evaluation of the model. It thereby prevents leakage of data from the test set [49]. Both LASSO and grid search should therefore ideally be performed using nested cross-validation.

3 Methods

3.1 Summary of methods

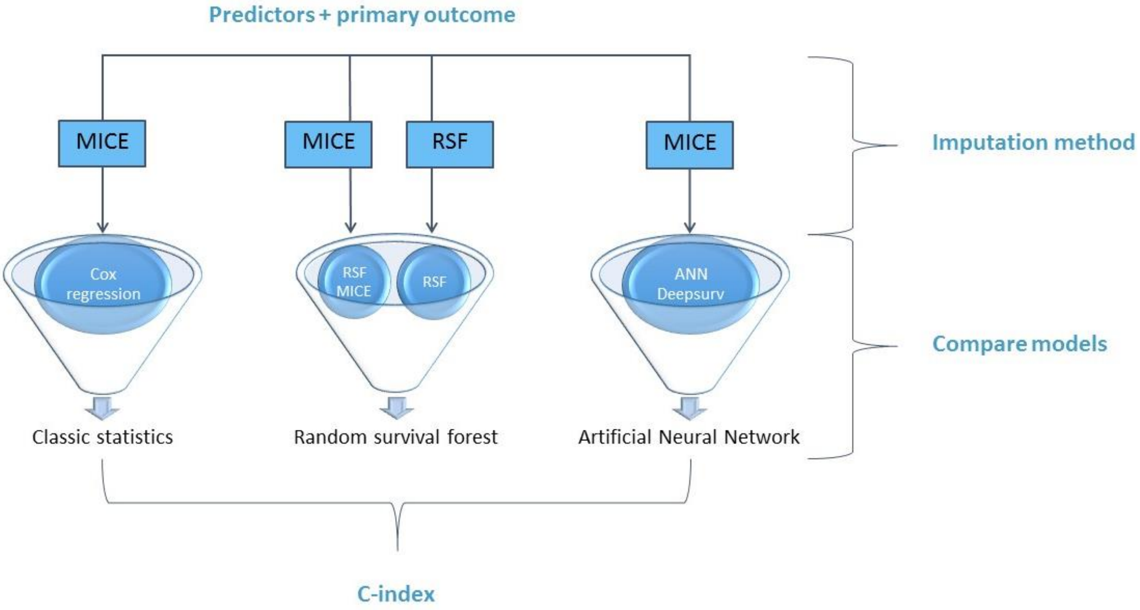


Figure 5: Summary of methods used in this research for every single dataset

Four different models (Cox regression, RSF-MICE, RSF-regular and ANN-DeepSurv) were constructed and compared regarding their c-indexes to assess the best fit. Missing data points were imputed using MICE before applying Cox regression, RSF-MICE and ANN-DeepSurv, while RSF-regular made use of its own imputation method. An overview is given in Figure 5. All models were trained and tested on two datasets: the complete dataset and the osteopenia subset. Detailed information regarding used methods are given in the next sections.

3.2 Study design

This retrospective cohort study was performed in the Ziekenhuisgroep Twente (ZGT), location Almelo. It is a non-WMO subject study as it comprises a retrospective data analysis study. The local ethics review committee of the ZGT gave their approval for this study (appendix A). All consecutive patients that sustained a (minor) fracture and visited the osteoporosis screening clinic of the ZGT between July 2011 and November 2019 were included in this research. An exclusion criterion was age < 50 as the Dutch guideline ‘Osteoporosis and fracture prevention’ recommends screening for patients of 50 years and older [12]. The primary endpoint of the study was the time till the occurrence of a MOF, defined in line with Briot et al. (2013) as humerus-, wrist-, spine- or hip fracture [3].

3.3 Data extraction

Data were extracted from the electronic health record using Structured Query Language (SQL) queries and were anonymised for analysis. Quality control was performed manually after every extraction for a random sample of 25 patients to ensure accuracy of the data.

Time till the occurrence of a MOF was extracted using diagnosis treatment combinations (DBCs in Dutch) as defined and labelled by the Dutch healthcare authority [50]. DBCs are healthcare products used for financial administration and therefore are a solid source of registration. The starting date for all patients was their visit to the osteoporosis screening clinic. For patients with a DBC regarding MOF following their visit to the osteoporosis screening clinic, the registration date of the DBC was used to calculate the total time till the occurrence of a MOF. Patients with no DBC of a MOF were considered censored. Either the date of death or, when alive, the date of the end of the study (15-11-2019) was used to calculate the follow-up time of censored patients.

DBCs and completed forms from the osteoporosis screening clinic were used to determine relevant comorbidities. History of a comorbidity and presence of a current comorbidity could not be distinguished as DBCs only return information on the start of the diagnosis and therefore were combined into one single variable. Both clinical parameters and parameters regarding lifestyle were extracted using osteoporosis screening clinic forms. Biochemical parameters were extracted using a time window of 6 months prior to the visit at the osteoporosis screening clinic until 1-week post visit.

3.3.1 Predictors

A literature study was performed to define risk factors for the occurrence of MOF. Databases consulted included Cochrane Central, Embase, MEDLINE, Pubmed, Scopus and Google Scholar. Search terms used were ‘major osteoporotic fracture’, ‘osteoporosis’, ‘subsequent fracture’ and ‘risk factors’, amongst others. Abstracts of reviews, systematic reviews and meta-analysis were read to assess the relevance of the article; all relevant articles were read to determine possible risk factors. The final set of predictors was selected based on expert opinion and retrospective availability.

Demographic study parameters included age and gender of the subject. Relevant comorbidities selected were use of corticosteroids, 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), ever being bedridden and a positive family history of a first-degree relative with either hip fracture or osteoporosis were considered for all patients. Number of children, use of combined oral contraceptive pill (COCP), breastfeeding to infants and duration of menopause were collected for all female patients. Clinical parameters collected included presence of vertebral fracture, reporting back pain, weight <67 kg, weight <60 kg, diminished length in recent years, length (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 (milligram), >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 D_3 and estimated glomerular filtration rate (eGFR[1]).

3.3.2 Data preparation

Variables which only apply to a subset of the patients were treated using two-part variables as described by Dziak and Henry (2017) [51]. A two-part variable is a method describing a variable as a pair of interrelated covariates. The first part is a dummy variable indicating if a covariate is relevant or not, while the second part gives the actual value if applicable. For all patients to which the variable does not apply, the actual value is set to zero [51].This two-part method was used for the variables number of children, use COCP, breastfeeding to infants and duration of menopause as these were only relevant for female patients. Besides, it was used for the eGFR as this variable contained both data points defined as ‘>90’ and numeric data points. Therefore, the eGFR was dichotomised in <90 mL/min/1.73m^2 and ≥90 mL/min/1.73m^2 as a dummy variable, describing a normal kidney function and renal insufficiency respectively. Interactions terms were created and included in the database by taking the product of two variables. Prior to creating the interaction term, continuous variables were first mean centred. Interaction terms with the variable age were created for gender, moderate active hours per week and T-scores of the hip and lumbar spine. For gender, interaction terms were created with the variables moderate active hours per week and the T-scores of the hip and lumbar spine. Outliers were identified using boxplots and were removed from the dataset.

[1] Determined with the CKD-EPI formula

3.3.3 Datasets

Two separate datasets were created for analysis in this study. The first dataset comprised all included patients as described earlier and will from now on be referred to as ‘complete dataset’. As patients with a T-score < -2.5 or a vertebral fracture are already treated with bisphosphonates, patients of interest for targeting preventive measures are those with osteopenia. Therefore, the second dataset used was a subset of the first dataset, containing only patients with osteopenia in the absence of a vertebral fracture. This dataset will be referred to as the ‘osteopenia subset’. All statistical procedures were performed for both datasets separately.

3.3.4 Imputation of missing data

The complete dataset had 2271 (30%) incomplete cases, while in the osteopenia subset 590 (33%) cases were incomplete. Data was imputed using the MICE package in R. In line with the advice of White and Royston (2009) [52], the Nelson-Aalen estimator of the cumulative baseline hazard and the event indicator were included in the imputation model. The predictor matrix for imputation was defined using the Quickpred function in the MICE package. The estimated hazard, the event indicator, gender and age were included as predictors for every variable, while time till event was excluded. The minimum threshold of absolute correlation (Pearson) was set to 0.1 and the minimum threshold for the proportion of usable cases to 0.5. This resulted in a mean of 30 predictors per variable in the complete dataset and a mean of 29 predictors per variable in the osteopenia subset, which is in line with the advice of Van Buuren and Groothuis-Oudshoorn (2011) to include at least 15 to 25 predictors [25]. The ratio between the original terms and the interactions terms was maintained by passive imputation using the ‘meth’ definitions in MICE. Numeric variables were imputed by predictive mean matching, while factor variables with two levels were imputed by logistic regression. As 30% (n=2271) of the cases were incomplete, the number of imputations used was set to 30. Fifty iterations were used, convergence was checked by plotting the mean of the synthetic data against the iteration number. The imputed data was checked using density plots to assess the model fit and possible distributional discrepancies. Furthermore, a stacked dataset of weighted observations was created to perform various statistical tests, as the pool function in MICE is not compatible with a couple of other functions. The stacked dataset was created by merging all 30 datasets with a weight of 1/30 for each patient in line with Steyerberg (2019) [20].

3.4 Statistics and used software

Descriptive statistics are provided for both cohorts and were compared using the chi-square test for nominal variables and Mann-Whitney U test for continuous variables. P-values <0.05 are considered statistically significant. Software used for data preparation and analysis in this research were R (R Core Team 2019) and PyCharm (Jetbrains 2019) [53,54].

3.4.1 Cox Proportional hazard regression model

The Cox proportional hazard regression model was constructed using the coxph function from the survival package in R. The assumption of proportional hazard was checked using the Schoenfeld's residual correlation test. The linearity assumption was tested using the likelihood ratio test and graphically checked by an effect plot of the log hazard ratio. Several transformations of continuous variables were considered and compared using the χ^2-statistic of the univariate Wald test. We considered the following transformations for variables x: log(x), $\log(x)^2$, x^2, $\sqrt{x}$, and restricted cubic splines of x with 3 or 4 degrees of freedom. Variables for the final model were selected using LASSO. In line with Steyerberg (2011), the penalty factor λ (or λ + 1 standard error) was determined in each imputed dataset using 10-fold cross validation and subsequently averaged. This penalty factor was applied to the stacked dataset to select the covariates for the final model. The final model was run for every imputed dataset, results were pooled using the pool function of MICE. The coefficients of the variables transformed by a restricted cubic spline were calculated as demonstrated by Shepherd and Rebeiro (2018) using the stacked dataset [55].

3.4.2 Random Survival Forest

Two RSF models were constructed in R. The first model was trained and validated on every single dataset created by the MICE algorithm, from now on referred to as RSF-MICE. The second model used the imputation data algorithm of the rsfrc function designed by Ishwaran et al. (2008) and will be referred to as RSF-regular. As both datasets were imbalanced, bs.gradient was used as a split rule. Variable importance was determined for both datasets. Due to incompatibility problems with the pool function of MICE, a single imputed dataset was used to determine variable importance for RSF-MICE.

3.4.3 Artificial neural network: DeepSurv

We used the pysurvival library to construct the DeepSurv model (ANN-DeepSurv) using python. As DeepSurv is not able to handle missing data itself, imputation of missing data is required. However, Python has not as many options as R for imputation of missing data. Recently, Kearney and Barkat (2019) started filling in this gap by designing a python package for multiple imputation called autoimpute. This package is currently compatible with linear regression and binary logistic regression but is not able to handle survival data nor ANNs [56]. For pragmatic reasons, we use the 30 completed datasets imputed by the MICE algorithm for the analysis using DeepSurv. Data was normalized prior to model development. Hyper-parameters of DeepSurv were optimized using grid search with 10-fold cross-validation on a single imputed dataset. Constraints for the hyper-parameters were set manually, weighing both computational time and model performance. The optimization algorithm used was the Adaptive Moment Estimation (Adam), while l2-regularization was set to default ($1e^{-4}$). The constraints used can be found in Table 1. Hyper-parameters of the model with the largest c-index were selected for final evaluation.

Table 1: An overview of the hyper-parameters of ANN-DeepSurv with corresponding constraints used for grid search.

Hyper-parameter	Constraints used for grid search (start/stop/step)
Learning rate	0.0001/0.01/0.001
Dropout	0.0/0.5/0.1
Number of nodes per layer	20/50/5
Number of layers	1/2/1
Activation function	1 layer: RELU or SELU 2 layers: RELU/RELU, RELU/SELU, SELU/RELU, SELU/SELU

Abbreviations: RELU = Rectified Linear Unit, SELU = Scaled Exponential Linear Unit

3.4.4 C-index, cross-validation and pooling

In this research, the c-index and its confidence interval were determined in R for the Cox regression model and the RSF models using 10-fold cross validation by the cindex function of the pec package. As this function is not compatible with the pool function of MICE, this procedure was performed for every individual dataset and eventually averaged for the Cox regression and RSF-MICE model. Averaging the results was performed with respect to Rubin’s rules using the pool.scalar function in R. For the ANN-DeepSurv, 10-fold cross validation was used for model evaluation in every imputed dataset in Python using the concordance_index function of the pysurvival library and the Kfold function of scikit-learn library. The c-index and its confidence interval were determined by averaging these results, again by using the pool.scalar function in R. The mean c-indexes were compared using an unpaired two-sample T-test (two-sided).

4 Results

In this study, 7578 patients were included, 5014 (74%) were female. In total, 805 (11%) patients sustained a subsequent MOF, while 6773 (89%) did not. Median time till event was 114 weeks (Interquartile range (IQR) = 224), while the median censored time was 192 weeks (IQR = 153). Median age for all patients was 68 (IQR = 17), and 74 (IQR = 15) and 67 (IQR = 16) for patients who sustained a subsequent MOF and censored patients, respectively. The osteopenia subset, a subset of the complete dataset, consisted of 1770 patients of which 1367 (77%) were female. In this subset, 165 (9%) patients sustained a subsequent MOF, while 1605 (91%) did not. Median time till event was 118 weeks (IQR = 147), while the median censored time was 159 (IQR = 217). Median age for all patients in the osteopenia subset was 67 (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 were lacking information, while for the osteopenia subset 590 (33%) cases were incomplete. The percentage of missing values across all variables varied between 0% and 23%. The primary endpoints, occurrence of a MOF and time till the occurrence of a MOF, were complete in both datasets. Percentages of missing data of each variable are given in Table 2. Density plots showed well matched distributions for the imputed and observed data for almost all variables in both datasets. Only the variables number of children and duration of menopause revealed a altered distribution for imputed and observed data at 0, as those variables were set to 0 for men. Density plots for all imputed continuous variables are attached in appendix B.

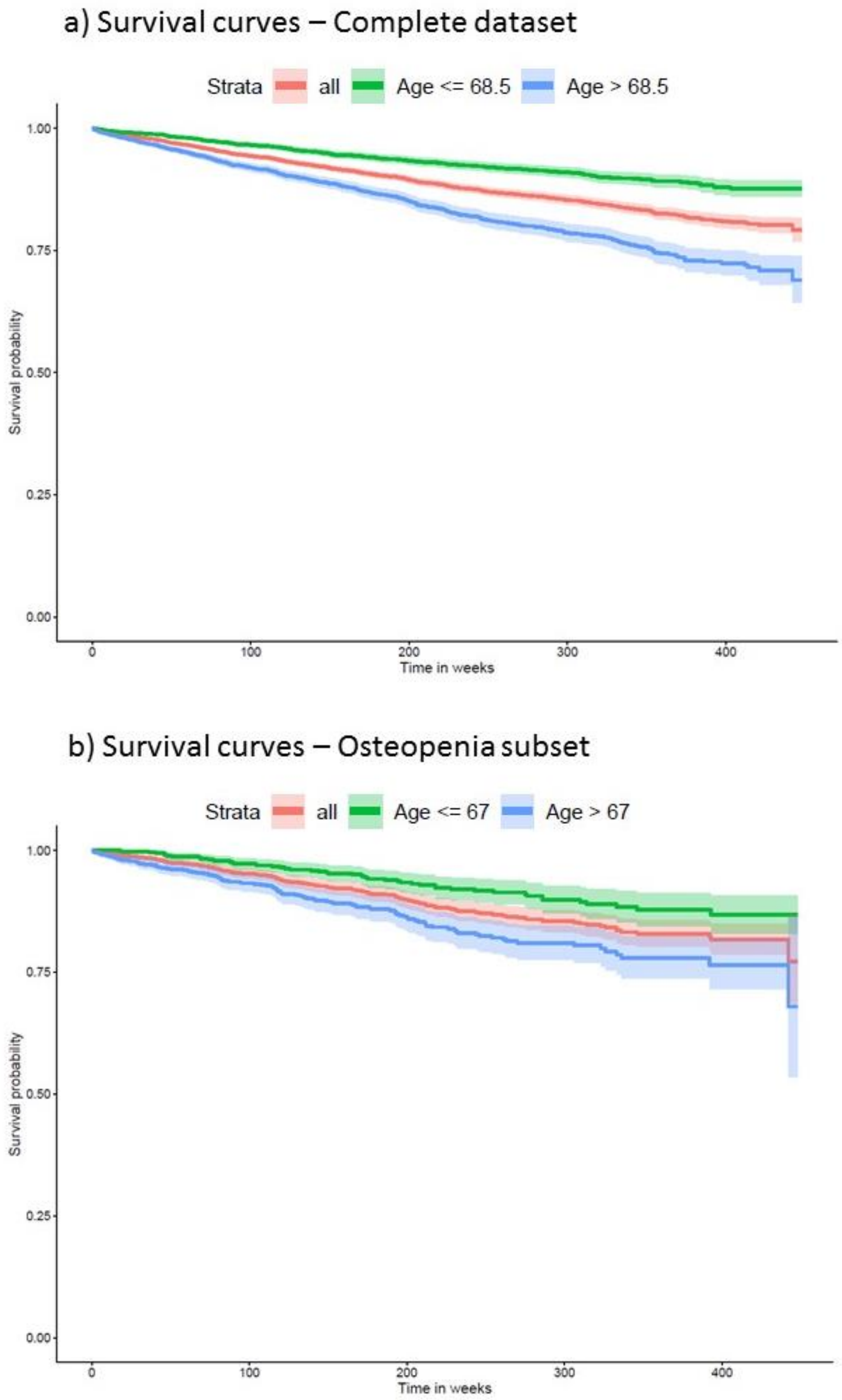
In the complete dataset, a significant correlation with MOF in univariate analysis was found in various covariates: age, gender, prior fall(s), current vertebral fracture, reporting back pain, number of children, use of COCP, breastfeeding to infant, duration of menopause, a history or presence of cardiovascular disease, epilepsy, rheumatoid arthritis, delirium or dementia, weight <67 kg and weight <60kg, change in length in recent years, moderate active hours per week, >6 cups of coffee per day, frequent exposure to sunlight, diet includes fat fish, ESR, plasma calcium, plasma albumin, serum vitamin D, eGFR, T-score of the lumbar spine and hip, weight and length. For the osteopenia subset, the variables age, prior fall(s), duration of menopause and change in length in recent years were significantly correlated with MOF. Descriptive statistics of both datasets stratified by MOF are shown in Table 2.

Table 2: Descriptive statistics for both the complete dataset (left) and the osteopenia dataset (right).

	Complete dataset				Osteopenia subset			
	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 (mL/min/1.73m^2)	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 (mm/h)	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 (g/l)	3.3%	38 [36, 40]	38 [35.50, 40]	<0.001	2.2%	38 [36, 40]	38 [36, 40]	0.878
Plasma calcium (mmol/l)	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 (mU/l)	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$_3$ (nmol/l)	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, n/a = not applicable 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. The survival rates over time are shown in Figure 6, stratified by the median of age.


a) Survival curves – Complete dataset

b) Survival curves – Osteopenia subset

Figure 6: Survival rates of the complete dataset (a) and osteopenia subset (b), stratified by the median of age with a 95% confidence interval

4.1 Cox regression

Continuous covariates age, plasma albumin and the T-scores for both lumbar spine and hip appeared linear with the log hazard and therefore met the linearity assumption of Cox proportional hazard regression. Other variables needed transformation to meet this assumption. The variables number of children and moderate active hours per week needed square root transformation, while ESR needed to be log squared. Duration of menopause, dietary daily calcium intake, plasma TSH and serum vitamin D were best fitted using restricted cubic splines with 3 degrees of freedom. Plasma calcium needed transformation using restricted cubic splines with 4 degrees of freedom. Cox proportional hazard assumptions were met (global chi-square of 53.3 (p-value 0.777) for the complete dataset and global chi-square of 59.9 (p-value 0.516) for the osteopenia subset). 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. This is shown in Figure 7. The minimum of λ was used in the osteopenia subset, instead of minimum λ plus 1 standard error, as the latter resulted in only one variable to be selected for final modelling. The c-index for the model selected with the minimum of λ, did not differ significantly from the model selected with minimum λ plus 1 standard error.

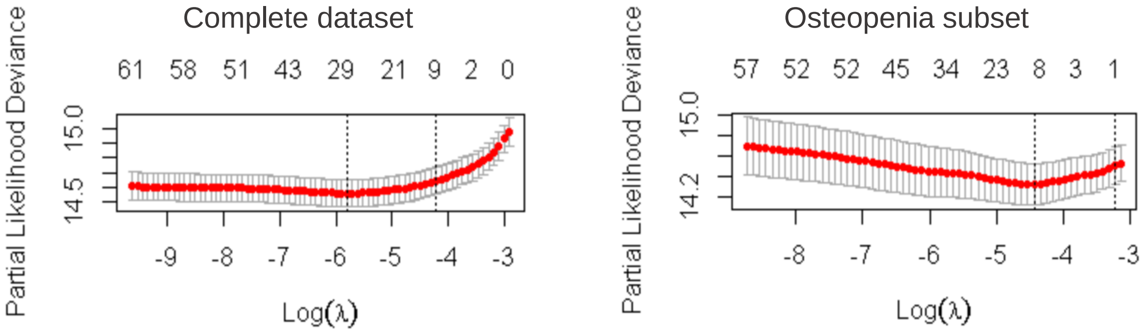


Figure 7: Illustration of the mean of the 10-fold Cross validation for the LASSO models in the complete dataset (left) and osteopenia subset (right), with the partial likelihood deviance on the y-axis and the natural logarithm of λ on the x-axis. The number above the graph describe the number of selected variables for logarithm of λ. The first dotted vertical line represents the minimum of λ, while the second dotted vertical line represents the minimum of lambda plus 1 standard error.

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. Duration of menopause was compared to the reference category of 10 years of menopause. 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), respectively). Results of the Cox regression in the complete dataset are shown in Table 3.

Table 3: Results of Cox proportional hazard regression of the 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

For patients in the osteopenia subset, prior fall(s), change in length in recent years, fat fish diet and renal insufficiency were significantly correlated with occurrence of MOF. The eGFR was compared to the reference category of 70 mL/min/1.73m^2. The hazard ratio increased for diminishing renal function. Results of the Cox regression of the osteopenia subset are shown in Table 4.

Table 4: Results of Cox proportional hazard regression of the 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
Renal insufficiency		3.218 (1.496-6.293)	0.003
eGFR (mL/min/1.73m^2)			<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

The Cox regression model returned a 10-fold cross validated c-index of 0.697 (0.664 – 0.730) for the total database and 0.625 (0.562 – 0.689) for the osteopenia subset.

4.2 RSF – MICE

As the total number of covariates is 46 and 45 for the different datasets, the RSF algorithm used seven variables as split candidates at every split. Cross validation (10-fold) 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 8 for both datasets.

4.3 RSF – regular

Like RSF-MICE, RSF-regular used seven variables as split candidates at every split. 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 10-fold 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. Again, the fifteen variables with largest variable importance are plotted in Figure 8 for both datasets.

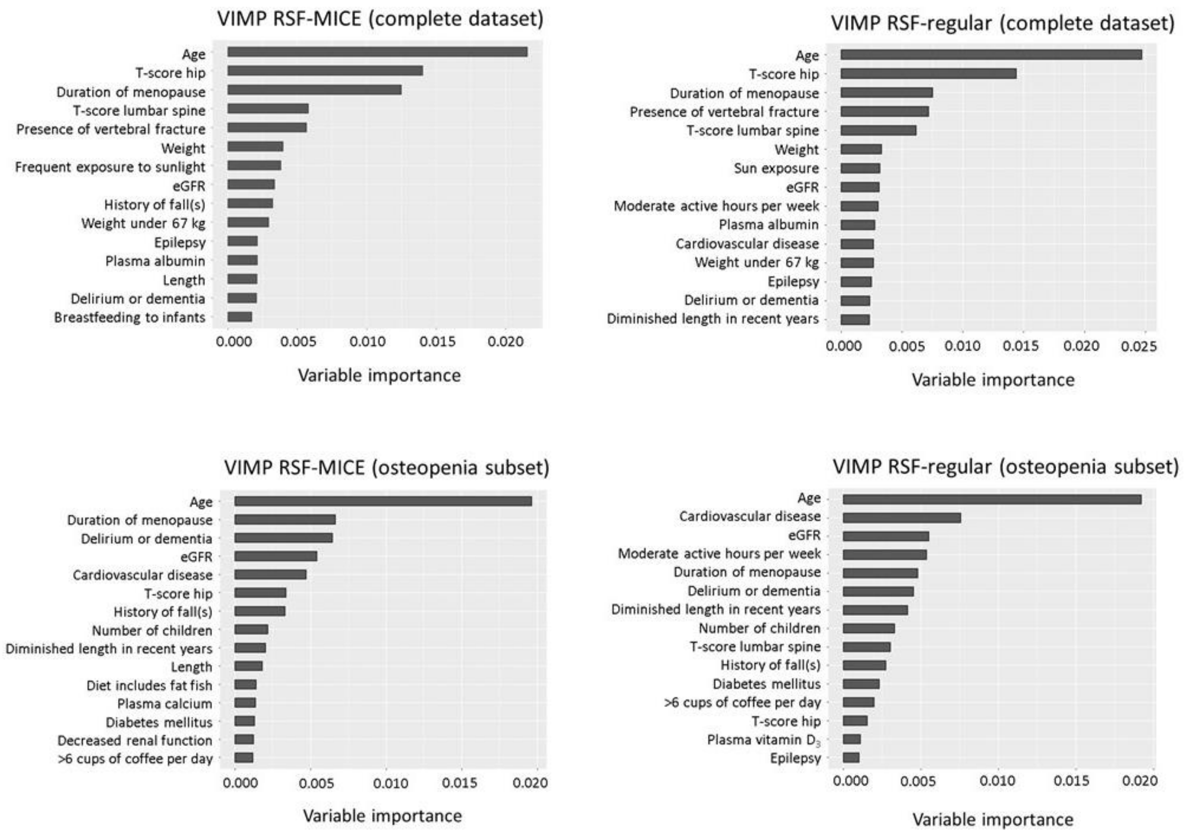


Figure 8: Variable importance of the fifteen variables with highest variable importance for every RSF model

4.4 ANN-DeepSurv

Optimal hyper-parameters as determined by grid search are shown in Table 5 for both datasets. As the ANN-DeepSurv model showed large variance, 10 final models were constructed per dataset and subsequently averaged. 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 10-fold cross-validation.

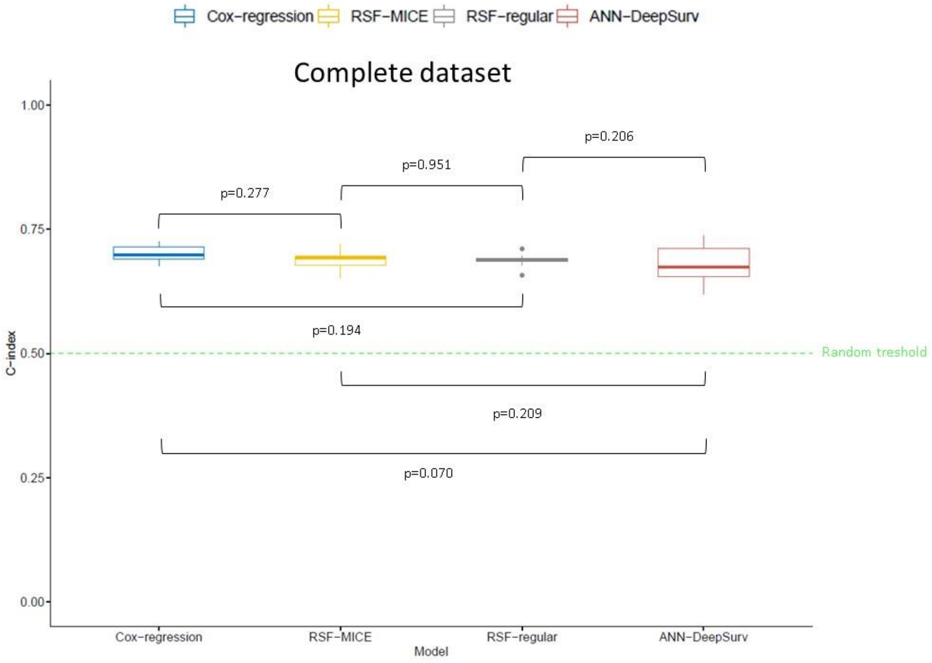
Table 5: Optimal hyper-parameter values in each dataset as determined using grid search

Hyper-parameter	Complete dataset	Osteopenia subset
Learning rate	0.0001	0.0001
Dropout	0.2	0.2
Number of nodes per layer	20 / 25	25
Number of layers	2	1
Activation function	RELU / SELU	RELU

Abbreviations: RELU = Rectified Linear Unit, SELU = Scaled Exponential Linear Unit

4.5 Model comparison

In 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=0.036), the RSF-regular model (p=0.026) and the ANN-DeepSurv model (p=0.042) on discriminative ability. Comparison of the c-indexes of the four models are given in Figure 9 for each dataset.



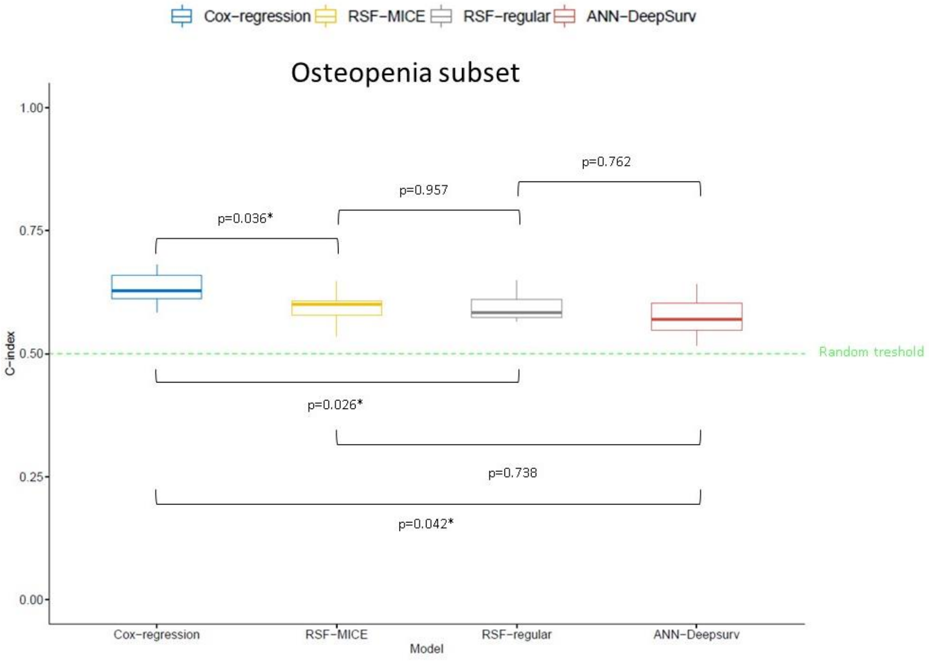
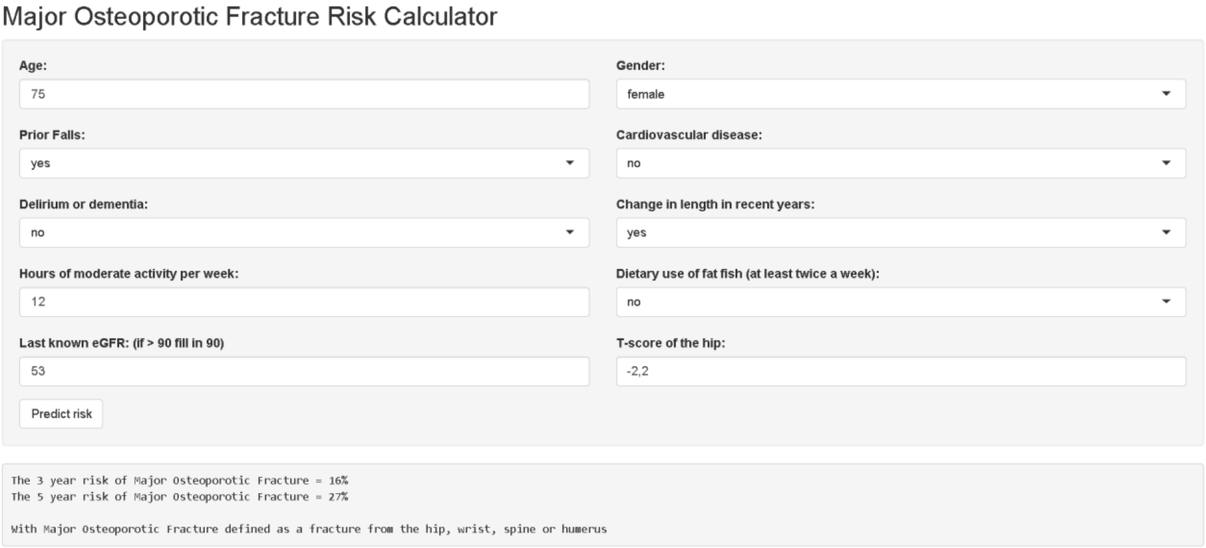


*Figure 9: Boxplot of the c-index of the predictive models: Cox regression, RSF-MICE, RSF-regular and ANN-DeepSurv. Corresponding p-value of the unpaired two-sample T-test for every individual comparison is shown. Sign * indicates statistical significance (p<0.05).*

4.6 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. An example of this risk calculator is shown in figure 10.



Major Osteoporotic Fracture Risk Calculator

Age: 75

Gender: female

Prior Falls: yes

Cardiovascular disease: no

Delirium or dementia: no

Change in length in recent years: yes

Hours of moderate activity per week: 12

Dietary use of fat fish (at least twice a week): no

Last known eGFR: (if > 90 fill in 90) 53

T-score of the hip: -2,2

```
The 3 year risk of Major Osteoporotic Fracture = 16%
The 5 year risk of Major Osteoporotic Fracture = 27%

With Major Osteoporotic Fracture defined as a fracture from the hip, wrist, spine or humerus
```

Figure 10: Major Osteoporotic Fracture Risk Calculator

5 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 already sustained a (minor) fracture. We developed four models that adequately predict the risk of MOF as a function of time in these patients and determined the predictive ability of Cox regression model, RSF models and ANN-DeepSurv model for patients at the Fracture and Osteoporosis Outpatient 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 patients with osteopenia at an Osteoporosis and Fracture Outpatient clinic.

Both 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 this research and returned significantly lower c-indexes in the osteopenia subset. This contrasts 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 [42]. Likewise, Kim et al. (2019) showed that RSF and DeepSurv outperformed Cox regression in survival prediction of oral cancer patients [57]. 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 these studies used 5 to 14 variables. Besides, our dataset was more imbalanced compared to the datasets of these studies. 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 this 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. 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 [42,57]. Therefore, the Cox regression models in their studies may not perform optimally. Besides, both studies only used single cross-validation to evaluate the model performance. Another remarkable finding in this research was the broad confidence interval of the c-index for the ANN-DeepSurv model compared to the other models. Again, the number of variables and the class-imbalance may play an important part, as Mazurowski et al (2007) showed that class-imbalance severely increases variability of performance in neural networks [58]. 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.

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 [57,58]. 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 [16,61]. 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 [62]. The incidence rate of 10.6% 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 Fracture and Osteoporosis Outpatient Clinic. All patients sustained a recent fracture and are therefore known to be at risk of a new fracture [63]. 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 model with a c-index of 0.69 (0.63-0.75), while Ensrud et al (2009) report a likewise c-index of 0.69 (0.67-0.70). These results are comparable to our findings, 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 [64]. Again, this is comparable to our results.

Moreover, we identified multiple risk factors of MOF for patients at the Fracture and Osteoporosis Outpatient Clinic, which may be useful in predictive modelling. Epilepsy, history of fall(s), presence of vertebral fracture and duration of menopause were determined to be independent risk factors for MOF, but are currently not used in the FRAX®-tool [65]. Frequent exposure to sunlight was determined as a protective factor in this study. For patients with osteopenia, history of fall(s), diminished length in recent years, dietary use of fat fish and decrease in renal function were all significantly associated with the risk of MOF. Additionally, cardiovascular comorbidity was determined as a contributing factor by the RSF model for osteopenia patients. Again, the FRAX®-tool does not take any of these factors into account.

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 [3,66,67,68]. 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 [69]. 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 [70]. Especially omega-3 fatty acids, which are high in fat fish, are known to have a protective effect on bone health and lower the risk of hip fracture in general population [71]. The negative effect of dietary use of fat fish in this study may be caused by specific advice given in our institute to patients with fractures, a diminished T-score or a lower serum vitamin D. We advise them to include fat fish in their diet to increase vitamin D level. This effect may therefore mimic a lower bone quality, as these patients may have received this advice in the years prior to inclusion in this study. However, this could not be verified as no data regarding this issue was available. The current vitamin D level of patients with dietary use of fat fish was significantly higher (54.4 vs. 51.6, p=0.02), while their T-score of the hip (-1.59 vs. -1.61, p=0.31) and lumbar spine (-1.76 vs. -1.77, p=0.65) did not differ. Therefore, the exact cause remains unknown.

At a Fracture and Osteoporosis Outpatient Clinic, risk identification for patients with osteopenia in absence of vertebral fracture is most relevant, as these patients are not standardly treated with bisphosphonates. 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 Fracture and Osteoporosis Outpatient 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. 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, 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 were not available due to the retrospective design of the studies. For example, 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 might be added in future research and may increase the discriminative ability of the models. We recommend adding dose and frequency of glucocorticoids use, number of falls, smoking, use of alcohol, Charlson comorbidity index and living situation in future studies. Besides, the treatment decision for every patient was not clearly recorded. 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 bisphosphonates while all others were not. Individual circumstances may however have led to different treatment decisions in the osteopenia subset and thereby trouble our results. Future research with a clear 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. Variable selection using LASSO, the interpretation of the effects of the restricted cubic splines and hyper-parameter optimization were therefore performed using a stacked dataset or a single imputed dataset. Especially the ANN-DeepSurv model may have suffered from this constraint, as the hyperparameter search was performed on a single dataset. Besides, we were technically not able to include nested-cross validation in both variable selection and hyper-parameter optimization. This might have led to leakage of test-data, as explained in the theoretical background. This could possibly result in too optimistic results for the models. 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. The occurrence of MOF might therefore be underestimated in our study. DBCs also do not distinguish between a history of a comorbidity and presence of a comorbidity. 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 conclusion, this pilot 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 that already sustained a (minor) fracture. Besides, we are the first in this field of research to combine both clinical and financial data for predictive modelling. We show adequate and comparable discriminative performance of a Cox regression model, RSF models and ANN-DeepSurv model in the population of a Fracture and Osteoporosis Outpatient 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 and initiated therapy, 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 risk of subsequent fracture and therefore help targeting treatment to patients at highest need.

6 References

[1] A. Odén, E. V. McCloskey, J. A. Kanis, N. C. Harvey, and H. Johansson, “Burden of high fracture probability worldwide: secular increases 2010–2040,” *Osteoporos. Int.*, vol. 26, no. 9, pp. 2243–2248, 2015.

[2] O. Johnell and J. A. Kanis, “An estimate of the worldwide prevalence and disability associated with osteoporotic fractures,” *Osteoporos. Int.*, vol. 17, no. 12, pp. 1726–1733, Oct. 2006.

[3] K. Briot *et al.*, “FRAX®: Prediction of Major Osteoporotic Fractures in Women from the General Population: The OPUS Study,” *PLoS One*, vol. 8, no. 12, p. e83436, Dec. 2013.

[4] M. Jürisson *et al.*, “Quality of life, resource use, and costs related to hip fracture in Estonia,” *Osteoporos. Int.*, vol. 27, no. 8, pp. 2555–2566, Aug. 2016.

[5] M. J. Goldacre, S. E. Roberts, and D. Yeates, “Mortality after admission to hospital with fractured neck of femur: database study.,” *BMJ*, vol. 325, no. 7369, pp. 868–9, Oct. 2002.

[6] E. Hernlund *et al.*, “Osteoporosis in the European Union: medical management, epidemiology and economic burden,” *Arch. Osteoporos.*, vol. 8, no. 1–2, p. 136, Dec. 2013.

[7] A. Singer *et al.*, “Burden of Illness for Osteoporotic Fractures Compared With Other Serious Diseases Among Postmenopausal Women in the United States,” *Mayo Clin. Proc.*, vol. 90, no. 1, pp. 53–62, Jan. 2015.

[8] R. Bouillon, P. Burckhardt, and C. Christiansen, “Consensus development conference: prophylaxis and treatment of osteoporosis.,” *Am. J. Med.*, vol. 90, no. 1, pp. 107–10, Jan. 1991.

[9] World Health Organization., “Assessment of fracture risk and its application to screening for postmenopausal osteoporosis : report of a WHO study group [meeting held in Rome from 22 to 25 June 1992].,” *World Health Organization*. 1994.

[10] M. Schurink, J. H. Hegeman, H. G. Kreeftenberg, and H. J. Ten Duis, “Follow-up for osteoporosis in older patients three years after a fracture.,” *Neth. J. Med.*, vol. 65, no. 2, pp. 71–4, Feb. 2007.

[11] A. H. Warriner, N. M. Patkar, H. Yun, and E. Delzell, “Minor, Major, Low-Trauma, and High-Trauma Fractures: What Are the Subsequent Fracture Risks and How Do They Vary?,” *Curr. Osteoporos. Rep.*, vol. 9, no. 3, pp. 122–128, Sep. 2011.

[12] “CBO Richtlijn Osteoporose en Fractuurpreventie-2011.” [Online]. Available: https://www.volksgezondheidenzorg.info/bestanden/documenten/cbo-richtlijn-osteoporose-en-fractuurpreventie-2011. [Accessed: 21-Oct-2019].

[13] M. Weda, M. E. Jansen, and R. A. A. Vonk, “Personalised medicine - Implementatie in de praktijk en data-infrastructuren,” 2017.

[14] “Personalised medicine | European Commission.” [Online]. Available: https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/personalised-medicine_en. [Accessed: 24-Dec-2019].

[15] G. El-Hajj Fuleihan, M. Chakhtoura, J. A. Cauley, and N. Chamoun, “Worldwide Fracture Prediction,” *J. Clin. Densitom.*, vol. 20, no. 3, pp. 397–424, Jul. 2017.

[16] W. J. Tseng, L. W. Hung, J. S. Shieh, M. F. Abbod, and J. Lin, “Hip fracture risk assessment: Artificial neural network outperforms conditional logistic regression in an age- and sex-matched case control study,” *BMC Musculoskelet. Disord.*, vol. 14, no. 1, p. 1, 2013.

[17] T. P. Ho-Le, J. R. Center, J. A. Eisman, T. V Nguyen, and H. T. Nguyen, “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.*, vol. 2017, pp. 4207–4210, 2017.

[18] U. Ferizi, S. Honig, and G. Chang, “Artificial intelligence, osteoporosis and fragility fractures,” *Curr. Opin. Rheumatol.*, vol. 31, no. 4, pp. 368–375, Jul. 2019.

[19] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, Sep. 2008.

[20] E. W. Steyerberg, *Clinical Prediction Models*. Cham: Springer International Publishing, 2019.

[21] S. Van Buuren, *Flexible Imputation of Missing Data*. Taylor & Francis Group, LLC, 2012.

[22] D. B. Rubin, “Inference and Missing Data,” *Biometrika*, vol. 63, no. 3, p. 581, Dec. 1976.

[23] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. 1987.

[24] S. Van Buuren and C. G. M. Groothuis-Oudshoorn, “Multivariate Imputation by Chained Equations: MICE V1.0 User’s manual,” *TNO report PG/VGZ/00.038*. pp. 1–39, 2000.

[25] S. van Buuren and K. Groothuis-Oudshoorn, “mice : Multivariate Imputation by Chained Equations in R,” *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[26] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[27] A. Hazra and N. Gogtay, “Biostatistics series module 9: Survival Analysis,” *Indian J. Dermatol.*, vol. 62, pp. 251–7, 2017.

[28] D. R. Cox, “Regression Models and Life-Tables,” *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, Mar. 1972.

[29] D. W. Hosmer, S. Lemeshow, and S. May, *Applied Survival Analysis*. 2008.

[30] D. Schoenfeld, “Partial Residuals for The Proportional Hazards Regression Model,” *Biometrika*, vol. 69, no. 1, p. 239, Apr. 1982.

[31] L. Karlsson, “An Evaluation of Methods for Assessing the Functional Form of Covariates in the Cox Model,” 2016.

[32] F. E. Harell Jr., *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis (second edition)*. 2015.

[33] B. Ghojogh and M. Crowley, “The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial,” pp. 1–23, May 2019.

[34] C. M. Stein, “Inadmissibility of the usual estimator of the mean of a multivariate normal distribution,” *Proc. Third Berkeley Symp. Math. Stat. Probab. Vol. 1 Contrib. to Theory Stat.*, pp. 197–206, 1956.

[35] F. J. W. M. Dankers, A. Traverso, L. Wee, and S. M. J. van Kuijk, “Prediction Modeling Methodology,” in *Fundamentals of Clinical Data Science*, Cham: Springer International Publishing, 2019, pp. 101–120.

[36] R. Tibshirani, “The lasso method for variable selection in the Cox model,” *Stat. Med.*, vol. 16, no. 4, pp. 385–95, Feb. 1997.

[37] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Introduction to Statistical learning*. 2017.

[38] L. Breiman, “Random forest,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[39] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[40] N. Shahid, T. Rappon, and W. Berta, “Applications of artificial neural networks in health care organizational decision-making: A scoping review,” *PLoS One*, vol. 14, no. 2, p. e0212356, Feb. 2019.

[41] F. Jiang *et al.*, “Artificial intelligence in healthcare: past, present and future,” *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017.

[42] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, “DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network,” *BMC Med. Res. Methodol.*, vol. 18, no. 1, p. 24, Dec. 2018.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[44] M. Claesen and B. De Moor, “Hyperparameter Search in Machine Learning,” pp. 10–14, Feb. 2015.

[45] R. Ghawi and J. Pfeffer, “Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity,” *Open Comput. Sci.*, vol. 9, no. 1, pp. 160–180, Jan. 2019.

[46] S. J. Caetano, G. Sonpavde, and G. R. Pond, “C-statistic: A brief explanation of its construction, interpretation and limitations.,” *Eur. J. Cancer*, vol. 90, no. 1–2, pp. 130–132, Jul. 2018.

[47] T. Therneau and E. Atkinson, “Concordance,” 2019. [Online]. Available: https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf. [Accessed: 06-Mar-2020].

[48] N. Balakrishnan and C. R. Rao, *Handbook of Statistics: Advances in Survival Analysis*, vol. 23. .

[49] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model

selection," *BMC Bioinformatics*, vol. 7, no. 91, 2006.

[50] "DIS open data." [Online]. Available: https://www.opendisdata.nl/. [Accessed: 07-Feb-2020].

[51] J. J. Dziak and K. L. Henry, "Two-Part Predictors in Regression Models," *Multivariate Behav. Res.*, vol. 52, no. 5, pp. 551–561, 2017.

[52] I. R. White and P. Royston, "Imputing missing covariate values for the Cox model," *Stat. Med.*, vol. 28, no. 15, pp. 1982–1998, Jul. 2009.

[53] "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/." .

[54] "PyCharm: the Python IDE for Professional Developers by JetBrains." [Online]. Available: https://www.jetbrains.com/pycharm/. [Accessed: 28-Feb-2020].

[55] B. E. Shepherd and P. F. Rebeiro, "Assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines," *J. Acquir. Immune Defic. Syndr.*, vol. 74, no. 3, pp. e60–e63, Mar. 2017.

[56] "autoimpute · PyPI." [Online]. Available: https://pypi.org/project/autoimpute/. [Accessed: 14-Feb-2020].

[57] D. W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, and H. J. Kim, "Deep learning-based survival prediction of oral cancer patients.," *Sci. Rep.*, vol. 9, no. 1, p. 6994, May 2019.

[58] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance.," *Neural Netw.*, vol. 21, no. 2–3, pp. 427–36, 2012.

[59] V. Forgetta *et al.*, "Machine Learning to Predict Osteoporotic Fracture Risk from Genotypes," *bioRxiv*, p. 413716, 2018.

[60] T. Nissinen, "Convolutional neural networks in osteoporotic fracture risk prediction using spine DXA images," no. March, 2019.

[61] C. Kruse, P. Eiken, and P. Vestergaard, "Machine Learning Principles Can Improve Hip Fracture Prediction," *Calcif. Tissue Int.*, vol. 100, no. 4, pp. 348–360, Apr. 2017.

[62] K. E. Ensrud *et al.*, "A comparison of prediction models for fractures in older women: is more better?," *Arch. Intern. Med.*, vol. 169, no. 22, pp. 2087–94, Dec. 2009.

[63] H. Johansson *et al.*, "Imminent risk of fracture after fracture Europe PMC Funders Group," *Osteoporos Int*, vol. 28, no. 3, pp. 775–780, 2017.

[64] K. C. Reber *et al.*, "Development of a risk assessment tool for osteoporotic fracture prevention: A claims data approach," *Bone*, vol. 110, pp. 170–176, 2018.

[65] "Frax® - Fracture Risk Assessment Tool." [Online]. Available: https://www.sheffield.ac.uk/FRAX/tool.aspx. [Accessed: 07-Mar-2020].

[66] M. Egan, S. Jaglal, K. Byrne, J. Wells, and P. Stolee, "Factors associated with a second hip fracture: a systematic review.," *Clin. Rehabil.*, vol. 22, no. 3, pp. 272–82, Mar. 2008.

[67] G. de Klerk, "Osteoporosis, identification and treatment in fracture patients," 2017.

[68] P. Haentjens, P. Autier, J. Collins, B. Velkeniers, D. Vanderschueren, and S. Boonen, "Colles fracture, spine fracture, and subsequent risk of hip fracture in men and women. A meta-analysis.," *J. Bone Joint Surg. Am.*, vol. 85, no. 10, pp. 1936–43, Oct. 2003.

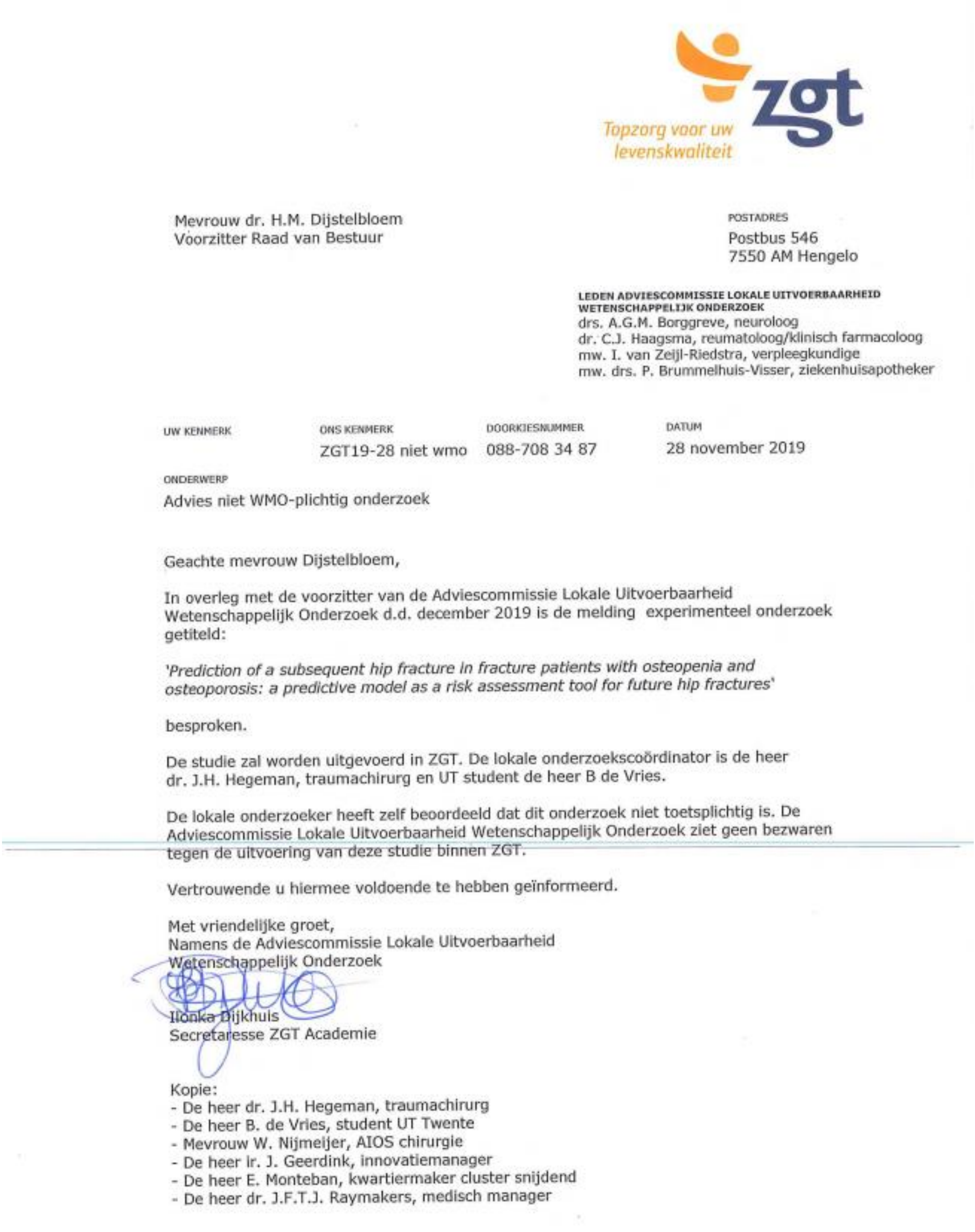
[69] D. Zhao, P. Cheng, and B. Zhu, "Epilepsy and fracture risk: A meta-analysis," *Int. J. Clin. Exp. Med.*, vol. 9, no. 2, pp. 564–569, 2016.

[70] S. Perna, I. Avanzato, M. Nichetti, G. D'Antona, M. Negro, and M. Rondanelli, "Association between dietary patterns of meat and fish consumption with bone mineral density or fracture risk: A systematic literature," *Nutrients*, vol. 9, no. 9, 2017.

[71] O. Sadeghi, K. Djafarian, S. Ghorabi, M. Khodadost, M. Nasiri, and S. Shab-Bidar, "Dietary intake of fish, n-3 polyunsaturated fatty acids and risk of hip fracture: A systematic review and meta-analysis on observational studies," *Crit. Rev. Food Sci. Nutr.*, vol. 59, no. 8, pp. 1320–1333, 2019.

7 Appendix

A) Approval of the local ethics review committee



Topzorg voor uw levenskwaliteit

zgt

Mevrouw dr. H.M. Dijstelbloem
Voorzitter Raad van Bestuur

POSTADRES
Postbus 546
7550 AM Hengelo

LEDEN ADVIESCOMMISSIE LOKALE UITVOERBAARHEID WETENSCHAPPELIJK ONDERZOEK
drs. A.G.M. Borggreve, neuroloog
dr. C.J. Haagsma, reumatoloog/klinisch farmacoloog
mw. I. van Zeijl-Riedstra, verpleegkundige
mw. drs. P. Brummelhuis-Visser, ziekenhuisapotheker

UW KENMERK	ONS KENMERK	DOORKIESNUMMER	DATUM
	ZGT19-28 niet wmo	088-708 34 87	28 november 2019

ONDERWERP
Advies niet WMO-plichtig onderzoek

Geachte mevrouw Dijstelbloem,

In overleg met de voorzitter van de Adviescommissie Lokale Uitvoerbaarheid Wetenschappelijk Onderzoek d.d. december 2019 is de melding experimenteel onderzoek getiteld:

'Prediction of a subsequent hip fracture in fracture patients with osteopenia and osteoporosis: a predictive model as a risk assessment tool for future hip fractures'

besproken.

De studie zal worden uitgevoerd in ZGT. De lokale onderzoekscoördinator is de heer dr. J.H. Hegeman, traumachirurg en UT student de heer B de Vries.

De lokale onderzoeker heeft zelf beoordeeld dat dit onderzoek niet toetsplichtig is. De Adviescommissie Lokale Uitvoerbaarheid Wetenschappelijk Onderzoek ziet geen bezwaren tegen de uitvoering van deze studie binnen ZGT.

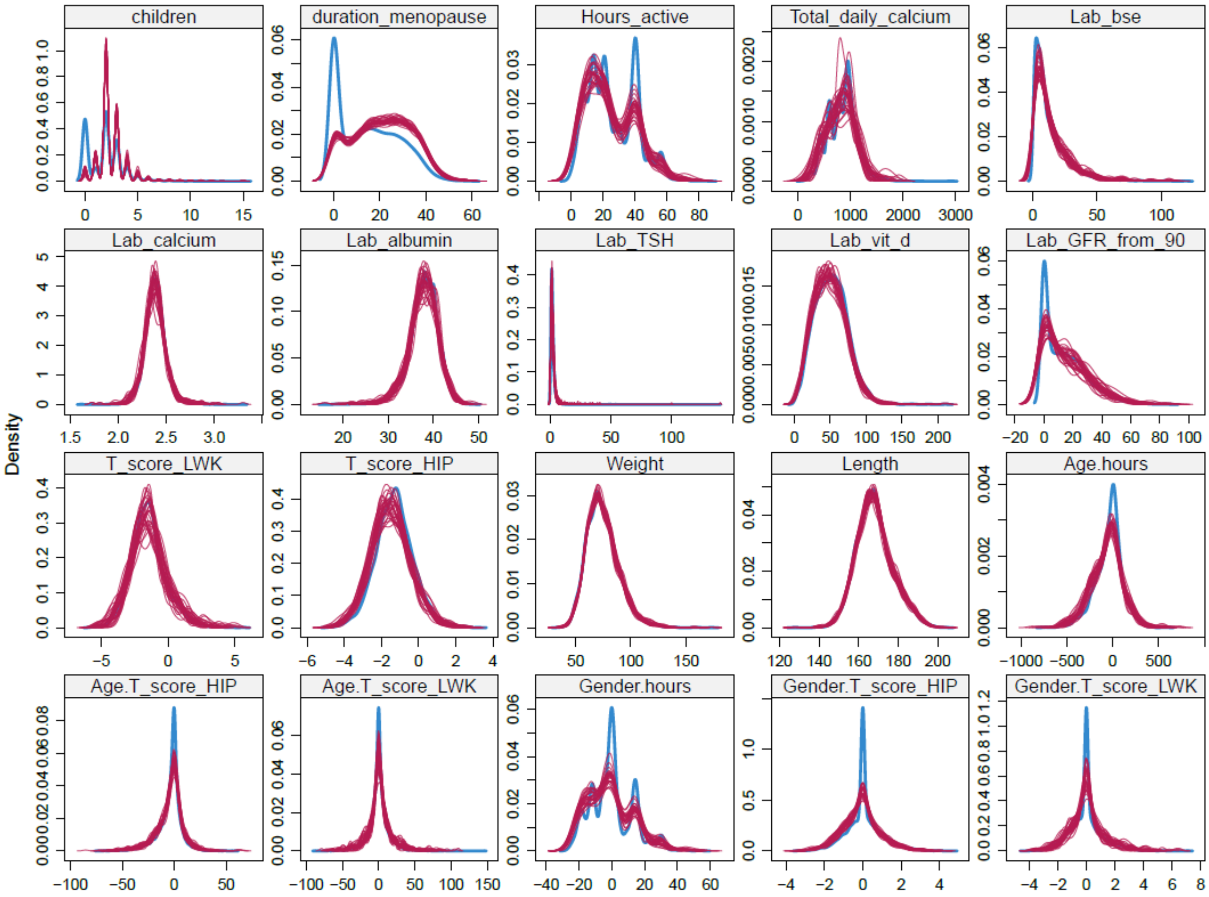
Vertrouwende u hiermee voldoende te hebben geïnformeerd.

Met vriendelijke groet,
Namens de Adviescommissie Lokale Uitvoerbaarheid
Wetenschappelijk Onderzoek

Ilonka Dijkhuis
Secretaresse ZGT Academie

Kopie:
- De heer dr. J.H. Hegeman, traumachirurg
- De heer B. de Vries, student UT Twente
- Mevrouw W. Nijmeijer, AIOS chirurgie
- De heer ir. J. Geerdink, innovatiemanager
- De heer E. Monteban, kwartiermaker cluster snijdend
- De heer dr. J.F.T.J. Raymakers, medisch manager

B) Density plots for all imputed continuous variables


Density
children
duration_menopause
Hours_active
Total_daily_calcium
Lab_bse
Lab_calcium
Lab_albumin
Lab_TSH
Lab_vit_d
Lab_GFR_from_90
T_score_LWK
T_score_HIP
Weight
Length
Age.hours
Age.T_score_HIP
Age.T_score_LWK
Gender.hours
Gender.T_score_HIP
Gender.T_score_LWK