[Category: Original Research]

Comparison of Machine Learning Approaches in Prediction of Bone Mineral Density in Elderly Men

Qing Wu, M.D., Sc.D (0000-0003-4679-8903)

Fatma Nasoz, Ph.D

Jongyun Jung, MS (0000-0002-9258-5397)

Bibek Bhattarai, MS

Author Affiliations:

Nevada Institute of Personalized Medicine (Dr Qing Wu and Jongyun Jung), University of Nevada, Las Vegas, Nevada;

Department of Environmental & Occupational Health, School of Public Health (Dr Qing Wu and Jongyun Jung), University of Nevada, Las Vegas, Nevada;

Department of Computer Science (Dr Fatma Nasoz and Bibek Bhattarai), University of Nevada, Las Vegas, Nevada;

The Lincy Institute (Dr Fatma Nasoz), University of Nevada, Las Vegas, Nevada;

Correspondence:

Qing Wu, MD, ScD, Nevada Institute of Personalized Medicine, Department of Environmental and Occupational Health, School of Public Health, University of Nevada Las Vegas, 4505 Maryland Parkway, Las Vegas, NV 89154-4009. Email: qing.wu@unlv.edu; Phone: 702-895-1439; Fax: 702-895-3094

Reprints:

Text word count: 3688

Abstract word count: 292

No. of tables: 5

No. of figures: 2

No. of supplemental tables: None

No. of supplemental figures: None

Running title: Machine learning for bone mineral density prediction

Abstract

Bone mineral density (BMD) is a highly heritable trait with heritability ranging from 50% to 80%. Numerous BMD-associated Single Nucleotide Polymorphisms (SNPs) were discovered by GWAS and GWAS meta-analysis. However, several studies found that combining these highly significant SNPs together only explained a small percentage of BMD variance. This inconsistency may be caused by limitations of the linear regression approaches employed because these traditional approaches lack the flexibility and the adequacy to model complex gene interactions and regulations. Hence, we developed various machine learning models of genomic data and ran experiments to identify the best machine learning model for BMD prediction at three different sites. We used genomic data of Osteoporotic Fractures in Men (MrOS) cohort Study (N=5,133) for analysis. Genotype imputation was conducted at the Sanger Imputation Server. A total of 1,103 BMD-associated SNPs were identified and corresponding weighted genetic risk scores were calculated. Genetic variants, as well as age and other traditional BMD predictors, were included for modeling. Data were normalized and were split into a training set (80%) and a test set (20%). BMD prediction models were built separately by random forest, gradient boosting, and neural network algorithms. Linear regression was used as a reference model. We applied the non-parametric Wilcoxon signed-rank tests for the measurement of MSE in each model for the pair-wise model comparison. We found that gradient boosting shows the lowest MSE for each BMD site and a prediction model built using the machine learning models achieves improved performance when a large number of SNPs are included in the models. With the predictors of *phenotype covariate + 1,103 SNPs*, all of the models were statistically significant except neural network vs. random forest at femoral neck BMD and gradient boosting vs. random forest at total hip BMD.

Keywords: Machine Learning, Bone Mineral Density, Osteoporosis, Genomics, Comparison

Abbreviations

MrOS : Osteoporotic Fractures in Men Study

BMD: Bone Mineral Density

ML: Machine Learning

GRS: Generic Risk Score

LR: Linear Regression

RF: Random Forest

GB: Gradient Boosting

NN: Neural Network

FNBMD: Femoral Neck BMD

TSBMD: Total Spine BMD

THBMD: Total Hip BMD

SNPs: Single Nucleotide Polymorphisms

GWAS: Genome Wide Association Study

Introduction

Osteoporosis is a major bone disease characterized by reduced bone mineral density (BMD) and deteriorated bone architecture, leading to increased fracture risk. Osteoporosis and its major complication, osteoporotic fracture, affecting both men and women, cause substantial morbidity and mortality worldwide[1]. Although women have a higher risk of osteoporosis, men suffer much higher morbidity and mortality rates following osteoporotic fractures, especially in their advanced age. With populations ageing worldwide, osteoporosis has become a critical public health problem globally. The worldwide fracture incidence in hip alone is projected to increase by 310% in men and 240% in women by 2050, compared to rates in 1990[2]. The potentially high cumulative rate of fracture, which often results in excess disability and mortality[3] has caused an inevitable increase in the social and economic burden associated with bone health.

BMD remains the operational definition of osteoporosis since 1994. Osteoporosis is defined by the World Health Organization as a BMD that lies 2.5 standard deviations or more below the average value for young healthy women. BMD is the single strongest predictor of primary osteoporotic fracture[4]. Each standard deviation decrease in BMD is associated with a 1.5-3.0 fold increase in the risk of fracture, depending on the skeletal region measured, type of fracture, and ethnicity of study population[5].

BMD is a highly heritable trait. Genetic differences in BMD are well documented[6]. Family and twin studies show BMD variances of 50-85% are attributable to genetic factors[7]. Other studies report BMD heritability estimates of 72-92[8]. In the past decade, major genome-wide association studies (GWASs) and genome-wide meta-analyses have successfully identified numerous BMD-associated Single Nucleotide Polymorphisms (SNPs) associated with decreased BMD[9]. However, combining these large number of highly significant SNPs together, surprisingly, only explained a very small percentage of BMD variance[10]. Such inconsistency may be caused by limitations of the conventional regression approaches employed because these conventional approaches lack the flexibility and the adequacy to model complex interactions and regulations. The Osteoporotic Fractures in Men Study (MrOS) is a unique prospective study in which BMD data and genotyping data are available.

Linear regression has been widely used as the traditional approach to predict BMD outcome[11]. ML focuses on implementing computer algorithms capable of maximizing predictive accuracy from complex data. ML has a much better capacity to model real-world complex relationships, including variable interactions. Several ML techniques have been applied in clinical research for disease prediction, and ML has shown much higher accuracy for diagnosis than conventional methods[12]. Gradient boosting, random forest, and neural network are widely used ML approaches for complex medical data[12]. However,

the performance of these ML models for BMD prediction remains unknown, especially with genomic data.

Hence, the aims of the current study are 1) to develop models using ML algorithms to predict BMD from the data with genomic variants; 2) to compare these models to determine which ML model performs the best for BMD prediction. We hypothesize that when we utilize the ML models to predict BMD, ML models will perform better than the linear regression (LR).

Materials and Methods

Data Source

The Osteoporotic Fractures in Men Study (MrOS) was used as the data source for this study. MrOS is a federal funded prospective, cohort study which was designed to investigate anthropometric, lifestyle, and medical factors associated with bone health in older, community dwelling men. Details of the MrOS study design, recruitment, and baseline cohort characteristics have been reported[12] elsewhere. With the approval of the institutional review board at the University of Nevada, Las Vegas and National Institute of Health (NIH), the genotype and phenotype data of MrOS were acquired from dbGaP (Accession: phs000373.v1.p1). MrOS have 5,130 subjects who have both genotype and phenotype data available for authorized access.

Study participants

Participants in the MrOS were at least 65 years old, community-dwelling, ambulatory, and had not bilateral hip replacement[13] at the study entry. At enrollment, participants had to provide self-reported data, understand and sign the written informed consent, complete the self-administered questionnaire (SAQ), attend a clinic visit, and complete at least the anthropometric, DEXA, and vertebral X-ray procedures. The participants could not have a medical condition that would result in imminent death based on the judgment of the investigators. A total 5,994 men were enrolled between March 2000 and April 2002 from six communities in the United States (Birmingham, AL; Minneapolis, MN; Palo Alto, CA; Pittsburgh, PA; Portland, OR; and San Diego, CA.)[14].

Outcome BMD measurements

Total body, total femur BMD, and lumbar spine (L1 to L4) were measured using a fan-beam dual-energy X-ray absorptiometry (QDR 4500 W, Hologic, Inc., Bedford, MA, USA) at the second visit of MrOS. Participants were scanned for BMD measurements by licensed densitometrists using standardized procedures. All DXA operators were centrally certified based on the evaluation results of scanning and analysis techniques. Cross-calibrations, which were conducted prior to participants' visits for BMD measurement, found no linear differences across canners, and the maximum percentage

difference between scanners was 1.4% in mean BMD of total spine[15]. No shifts or drifts in scanner performance was found based on longitudinal quality control using daily scan data for standardized phantoms in each clinical center.

We predict three outcomes: Femoral Neck BMD ($n = 5{,}129$), Total Spine BMD ($n = 5{,}120$), and Total Hip BMD ($n = 5{,}129$). We predict each BMD with two sets of predictors: (i) *phenotype covariates + GRS;* and (ii) *phenotype covariates + 1,103 SNPs*. We are also interested in the use of individual SNPs that are found by Morris et al.[9]. We encoded each 1,103 risk SNPs as three different genotypes (homozygous major allele, heterozygotes, homozygous minor allele) with 0, 1, and 2, respectively and then we treated it as a continuous variable to predict each BMD.

Assessment of covariates

Bone health related information, including demographics, clinical history, medications, and lifestyle factors were obtained by self-administered questionnaires. This information comprised variables used in this study, including age, race, smoking, and alcohol consumption. Height (cm) was measured using a Harpenden stadiometer and weight (kg) was measured by a standard balance beam or an electric scale. BMI was calculated as kilograms per square meter.

Smoking was categorized as “never”, “past” and “current”. Alcohol intake was quantified in terms of usual drinks per day. Walking speed was determined by timing completion of a 6-m course performed at the participant’s usual walking speed. Mobility limitations were measured by participants’ ability to rise from a chair without using their arms, as well as their ability to complete five chair stands. Health status was assessed as impairment in instrumental activities of daily living (IADLs).

Genotyping Data

Baseline whole blood samples were used for DNA extraction. Consent for use of DNA was obtained through written consent. Quality-control Plink genotype data files were acquired through dbGaP. Genotype imputation was conducted at the Sanger Imputation Server. Haplotype Reference Consortium (HRC) reference panel, the most comprehensive imputation reference panel, as well as Positional Burrows-Wheeler Transform (PBWT) imputing algorithm, were used for the genotype imputation to ensure high quality of genotype imputation. The most up-to-date GWAS study in the field, published by Morris et al in 2019, identified 1,103 SNPs associated with estimated BMD (eBMD)[9]. Each of the 1,103 SNPs were successfully imputed and were included in the analysis. The imputation quality was excellent with a mean R^2 of 0.99. Genotyping for MrOS samples was performed with the Illumina HumanOmni1_Quad_v1-0 H array. A total of 934,940 SNP markers with known chromosome locations, and SNP markers with minor allele frequencies greater than or equal to 0.05 were analyzed.

Genetic risk score

A genetic risk score (GRS) is a standardized metric derived from the number of risk alleles and their effect size for each study subject. This metric allows the composite assessment of genetic risk in complex traits. A linkage disequilibrium (LD) pruning was performed in advance to eliminate possible LD between SNPs. None of the 1,103 SNPs was removed after the pruning. The weighted GRS was then calculated with algorithms described previously[16]. Briefly, for each participant in MrOS, overall weighted GRS was calculated by summing the number of risk alleles at each locus multiplied by their effect size, which are regression coefficients related to BMD, from the referenced literature[6].

Model evaluation and validation

To evaluate the accuracy of the model, we randomly split a total of 5,130 patients into two parts: (i) the training set (80%), which was used to build the model, (ii) the test set (20%), which was used to evaluate the performance (i.e., the predictive accuracy) of the developed model. Commonly used metrics to evaluate the model performance are mean squared error and mean absolute error[17]. We adopted these two metrics:

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n is the sample size, y_i is the actual value for each observation, and $\hat{y}_i$ is the estimated value for each observation in the test dataset.

Hyper-parameter Optimization

10-fold cross-validation was used for hyper-parameter optimization. In 10-fold cross-validation, we chose one fold as a validation set and remaining folds as the training set. We used Scikit-learn's randomized search cross-validation method[18] to find the best hyperparameters for different algorithms. The training set was used to train and construct the models of logistic regression, random forest, gradient boosting, and neural network. Once the model was developed on the training set, then each model evaluated in a test set. We used a small learning rate and relatively small depth for RF and GB algorithms to reduce the risk of overfitting as it has been suggested that RF and GB are robust to overfitting[19]. With *phenotype covariate + GRS,* we used the depth 3 at THBMD, depth 2 at FNBMD and depth 2 at TSBMD. With *phenotype covariate + 1,103 SNPs,* we used the depth 4 at THBMD, depth 3 at FNBMD and depth

3 at TSBMD. Regularization is also a method to avoid overfitting[20,21] in the regression models. We used the Lasso[22] to avoid overfitting with *phenotype covariate + 1,103 SNPs* predictors in NN. In our study, statistical significance was set at $p < 0.05$ (5%).

Statistical Testing

Several studies suggest the use of statistical tests in machine learning for comparisons of many algorithms[23–25]. The most frequently used statistical tests to determine significant differences between two machine learning algorithms are the t-test and Wilcoxon signed-rank test[26]. The t-test is a parametric test that requires certain assumptions of independence, normality, and heteroscedasticity, which are usually not satisfied in the machine learning models. Thus, the nonparametric test of Wilcoxon signed-rank test is widely applicable to compare ML models.

Software used

All of the analyses were performed in the Python Software Foundation and Python Language Reference, version 3.7.3 with the package Scikit-learn: Machine Learning in Python[18]. Available at http://www.python.org

Results

Data analysis

Figure 1 shows the overview of our data process flow for this study. With two different MrOS data sets (Phenotype set ($n = 5{,}143$) and Genotype set ($n = 5{,}130$)), we first imputed the missing value for each data set as we discussed in the methods section. Next, we combined these two data sets to one data set for further analysis. When combining these two data sets, data from 13 patients in the MrOS Genotypes were removed due to their phenotype variables not being available in MrOS Phenotype data set. This resulted in one data set of ($n = 5{,}130$). Then we normalized the continuous predictors to have the same scale without distorting differences in the range of values. Normalization is a technique that is applied as a part of data preparation for machine learning models. We randomly split the dataset into a training set (80%, $n = 4{,}104$) and a test set (20%, $n = 1{,}026$). Outliers were not removed due to their relevance in the models. The most common method of missing data replacement is to use the median[27]. The median imputation was applied to fill the missing value in the variable.

Baseline characteristics

Table 1 shows the characteristics of participants within the training ($n = 4{,}104$) and the test ($n = 1{,}026$) datasets. The variables are not significantly different in training and test datasets. Each outcome variable of FNBMD, TSBMD, and THBMD follow the normal distribution with means 0.78, 1.07, 0.96 and standard deviations 0.13, 0.14, 0.19, respectively.

Model Performance

Figure 2 shows the performance of each model in the test dataset ($n = 1{,}026$). The first row (PC + GRS) of A, B, and C in Figure 2 shows the performance of each model with the predictor of *phenotype covariates + GRS* in the test dataset with MSE. Although the LR model had the relatively higher MSE in the first few iterations, LR and ML models are performing nearly identical with 100 iterations for each BMD site. Among these three BMD sites, FNBMD (A, in Figure 2) shows the lowest MSE, followed by THBMD (B, in Figure 2) and TSBMD (C, in Figure 2) in all of the models. The second row (PC + 1,103 SNPs) of D, E, and F show the performance of each model with the predictor of *phenotype covariates + 1,103 SNPs* in the test dataset of MSE. LR model results showed a higher MSE than ML models at each BMD site.

Tables 2 and 3 show the results of MSE and MAE for each model in the test dataset ($n = 1{,}026$) with two different sets of predictors *phenotype covariate + GRS* and *phenotype covariates + 1,103 SNPs*. With *phenotype covariate + GRS* (Table 2), there is no significant differences between linear regression and machine learning models. TSBMD has the highest MSE (GB = 0.0349), followed by THBMD (GB = 0.0162) and FNBMD (GB = 0.0142). With *phenotype covariates + 1,103 SNPs* (Table 3), ML models show the lower MSE and MAE in THBMD and TSBMD. Neural Network has the lowest MSE (0.0155) in the test dataset of THBMD and Gradient Boosting has the lowest MSE (0.0328, 0.0142) in the test datasets of TSBMD, FNBMD, respectively. TSBMD has the highest MSE (GB = 0.0328), followed by THBMD (NN = 0.0155) and FNBMD (GB = 0.0142).

The nonparametric Wilcoxon signed-rank tests were implemented for the measurement of MSE. The hypothesis for our testing is two models are equal, otherwise are not. The results are shown in Tables 4 and 5 for two different sets of predictors, *phenotype covariates + GRS* and *phenotype covariates + 1,103 SNPs*. In each cell results for a given pair of model was placed (i.e., p-value is placed to compare two models like Neural Network vs. Linear Regression). With Bonferroni corrections for multiple comparisons ($\alpha = 0.05/6=0.0083$), our results suggest that none of the models were statistically significant with *phenotype covariates + GRS* except gradient boosting vs. random forest at FNBMD and TSBMD. However, the difference of accuracy between two models in all pairwise comparisons was statistically significant with $p < .0001$, except Neural Network vs. Random Forest at FNBMD and Gradient Boosting vs. Random Forest at THBMD with $p > .05$.

Discussion

In this study, we investigated whether ML models perform better to predict the major BMD in MrOS data. It has been known that the performance of NN models are superior to predict BMD values

using several parameters among postmenopausal women than the conventional regression methods[11,28]. To the best of our knowledge, this is the first attempt to predict BMD at a different site, using various ML models by also including the genetic factors in Men's data. Machine learning employs a wide-ranging class of algorithms widely used to solve complex prediction problems. It has been very popular when prediction is dependent on the integration of a large number of predictors, including higher-order interactions, and when sizeable training datasets are available for model fitting. In particular, gradient boosting is powerful for continuous outcome prediction[29]. ML models have been used widely in classification problems (e.g., disease or not, fracture or not). However, there are only a few studies that applied ML models to predict the quantitative variable in the clinical study. Our models' performance results support that gradient boosting shows the lowest MSE for each BMD site, and a prediction model built using the ML models achieves improved performance when a large number of SNPs are included in the models[30].

We predicted each BMD site with two sets of predictors: 1) *phenotype covariate + GRS* and 2) *phenotype covariate + 1,103 SNPs* to predict BMD using various ML models. Our results showed that when we used the predictors of *phenotype covariate + GRS*, ML models are performing relatively similarly to LR, however, when we used the predictors of *phenotype covariate + 1,103 SNPs*, the result of model performance was consistent with our hypothesis that ML models are slightly better than LR. We assume that the low performance of ML with *phenotype covariate + GRS*, is due to the lack of predictors in the model. Since we only have 10 predictors including GRS, ML models are performing close to LR model. Since the use of ML approaches are suggested with the interaction of predictors[30], our selected ML models are performing better with predictors of *phenotype covariate + 1,103 SNPs.*

We applied the nonparametric Wilcoxon signed-rank tests, which were implemented for the measurement of MSE in each model for the pair-wise model comparison. With Bonferroni corrections for multiple comparisons, with the predictors of *phenotype covariate + GRS*, none of the models were statistically significant except gradient boosting vs. random forest at FNBMD and TSBMD. However, with the predictors of *phenotype covariate + 1,103 SNPs*, all of the models were statistically significant except Neural Network vs. Random Forest at FNBMD and Gradient Boosting vs. Random Forest at THBMD.

We anticipate several potential applications of our findings to novel types of clinical sites. The first application would be using the ML models to predict the BMD at different sites. Many studies found that the lower BMD is associated with higher fracture risk or osteoporosis disease[31]. Thus, ML models will help to detect the low BMD early. The second application would be using the risk SNPs to predict the BMD instead of the GRS as one variable. As our study results show, when we use the GRS as one of

the predictors variable, the prediction performance of ML models are relatively close with the linear regression. However, when we use the risk 1,103 SNPs with other covariates, the prediction results with ML models are better than the linear regression. We assume that since the ML models are performing better with more complex data[29], our result shows that the risk 1,103 SNPs with other covariates are performing better than the GRS with other covariates. Since our data set is relatively simple to utilize the ML models, our approaches of ML were feasible with depth equal to 2 or 3 in RF and GB models.

We have several limitations in our study. First, the total sample size ($n = 5{,}130$) is relatively small for ML approaches. ML approaches often require much larger training data size. For this reason, we employed 10-fold cross-validation for tuning of the hyper-parameters within the training set, instead of allocating part of the study sample for validation purpose, which would cause a smaller sample size for ML model training. Second, some covariates are not available in the MrOS through dbGaP, including medications, comorbidities, and physical activities. It may be plausible to conclude that the lack of these predictor variables in *phenotype covariate + GRS,* caused the ML models to perform similarly to the LR model. Third, the MrOS data we used only includes male participants who were ≥ 65 years old and mostly white (90%), so our findings may not apply to women or to individuals who are of a younger age and other ethnicity. Finally, we calculated GRS based on the common variants that were found by the Morris et al. study[9] and used these risk SNPs as our predictors along with other covariates. Recently this study[31] identified 1,362 independent SNPs, which are associated with BMD and fracture. Rare variants, which provide evidence that low-frequency non-coding variants have larger effects on BMD and fracture[32], were not included in our risk SNPs as predictors.

Conclusion

When compared with linear models, the use of a machine-learning approach, which models data non-linearity, could lead to an improvement in the accuracy of predicting each BMD with the complex data. Therefore, our findings can also be applied to other disease or women's data.

As with other prediction problems involving machine learning techniques, incremental improvements are to be expected with increased sample size, the inclusion of more predictors, and the availability of more precise summary association statistics to calculate GRS. With the ML models performing better than LR with the predictors of *phenotype covariate + 1,103 SNPs,* THBMD and TSBMD give slightly lower MSE and MAE. Early prediction of BMD is essential to diagnose the patient properly. Our ML models with the predictors of *phenotype covariate + 1,103 SNPs,* will help the

clinician's decision to treat the patient properly. Since the low BMD is the major risk factor for fracture in osteoporosis[33], accurate prediction of BMD is an essential step for the osteoporosis disease. We hope that ML approaches, with accurate prediction of low BMD along with other risk factors covariates as well as genetic factors, will help the clinician's care system with a risk patient.

Acknowledgments

The data/analyses presented in the current publication are based on the use of study data downloaded from the dbGaP web site, under phs000373.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000373.v1.p1). The research and analysis described in the current publication was supported by a COBRE grant from the National Institute of General Medical Sciences (5 U54 GM 104944), the Genome Acquisition to Analytics (GAA) Research Core of the Personalized Medicine Center of Biomedical Research Excellence in the Nevada Institute of Personalized Medicine, and the National Supercomputing Institute at the University of Nevada Las Vegas. The funding sponsor was not involved in the analysis design, genotype imputation, data analysis, and interpretation of the analysis results or in the preparation, review, and/or approval of this manuscript.

Compliance with ethical standards

Conflicts of interest

None.

Authors' roles

Study design: QW. Study conduct: QW. Data collection: JJ and BB. Data analysis: JJ and BB. Data interpretation: QW, FN, JJ, and BB. Drafting manuscript: QW and JJ. Revising manuscript content: QW, FN and JJ. Approving final version of manuscript: QW, FN, JJ, and BB. QW takes responsibility for the integrity of the data analysis.

References

1. Cummings, S. R. & Melton, L. J. Epidemiology and outcomes of osteoporotic fractures. *Lancet* **359**, 1761–1767 (2002).

2. Gullberg, B., Johnell, O. & Kanis, J. A. World-wide projections for hip fracture. *Osteoporos Int* **7**, 407–413 (1997).

3. Melton, L. J. & Cooper, C. Chapter 21 - Magnitude and Impact of Osteoporosis and Fractures. in *Osteoporosis* 557–567 (Academic Press Inc., 2007). doi:10.1016/B978-012470862-4/50022-2

4. Kanis, J. A. *et al.* Assessment of fracture risk. *Osteoporos Int* **16**, 581–589 (2005).

5. Marshall, D. & Wedel, H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *theBMJ* **312**, 1254–1259 (1996).

6. Warrington, N. M., Kemp, J. P., Tilling, K., Tobias, J. H. & Evans, D. M. Genetic variants in adult bone mineral density and fracture risk genes are associated with the rate of bone mineral density acquisition in adolescence. *Hum Mol Genet* **24**, 4158–4166 (2015).

7. Eisman, J. A. Genetics of Osteoporosis. *Endocr Rev* **20**, 788–804 (1999).

8. Nicholas A. Pocock, John A. Eisman, John L. Hopper, Michael G. Yeates, Philip N. Sambrook, S. E. Genetic determinants of bone mass in adults . A twin study. *Jounral Clin Investig* **80**, 706–710 (1987).

9. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat Genet* **51**, 258–266 (2019).

10. Xiao, X., Roohani, D. & Wu, Q. Genetic profiling of decreased bone mineral density in an independent sample of Caucasian women. *Osteoporos Int* **29**, 1807–1814 (2018).

11. Shioji, M. *et al.* Artificial neural networks to predict future bone mineral density and bone loss rate in Japanese postmenopausal women. *BMC Res Notes* **10**, 1–5 (2017).

12. Orwoll, E. *et al.* Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study - A large observational study of the determinants of fracture in older men. *Contemp Clin Trials* **26**, 569–585 (2005).

13. Riggs, L. & Melton, L. The worldwide problem of osteoporosis: lessons from epidemiology. *Bone* **17**, 2–3 (1995).

14. Blank, J. B. *et al.* Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemp Clin Trials* **26**, 557–568 (2005).

15. Cauley, J. A. *et al.* Factors associated with the lumbar spine and proximal femur bone mineral density in older men. *Osteoporos Int* **16**, 1525–1537 (2005).

16. Andrews, N. A. Genome-wide association studies in the osteoporosis field: Impressive technological achievements, but an uncertain future in the clinical setting. *IBMS Bonekey* **7**, 382–387 (2010).

17. Amoroso, N. *et al.* Deep learning and multiplex networks for accurate modeling of brain age.

Front Aging Neurosci **11**, 1–12 (2019).

18. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertran Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, D. C. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).

19. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann Stat* **28**, 337–407 (2000).

20. Yun Liu, Po-Hsuan Cameron Chen, Jonathan Krause, L. P. How to read articles that use machine learning. *JAMA - J Am Med Assoc* **322**, 1806–1816 (2019).

21. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning*. **103**, (2013).

22. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B* **58**, 267–288 (1996).

23. Demˇsar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J Mach Learn Res* **7**, 1–30 (2006).

24. García, S. & Herrera, F. An extension on ‘statistical comparisons of classifiers over multiple data sets’ for all pairwise comparisons. *J Mach Learn Res* **9**, 2677–2694 (2008).

25. García, S., Fernández, A., Luengo, J. & Herrera, F. A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Softw Comput* **13**, 959–977 (2009).

26. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bull* **1**, 80–83 (1945).

27. Gao, B. *Advances in Intelligent Systems and Computing*. *Springer* **997**, (2019).

28. M. Sadatsafavi, A. Moayyeri, A. Soltani, B. Larijani, M. Nouraie, S. A. Artificial neural networks in prediction of bone density among post-menopausal women. *J Endocrino Invest* **28**, 425–431 (2005).

29. Trevor Hastie, Robert Tibshirani, J. F. *The Element of Statistical Learning Data Mining, Inference, and Prediction*. *Springer* (Springer, 2009).

30. Wei, Z. *et al.* From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet* **5**, (2009).

31. Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS One* **13**, e0200785 (2018).

32. Zheng, H. F. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* (2015). doi:10.1038/nature14878

33. LaFleur, J., McAdam-Marx, C., Kirkness, C. & Brixner, D. I. Clinical risk factors for fracture in postmenopausal osteoporotic women: A review of the recent literature. *Ann Pharmacother* **42**, 375–386 (2008).

Table 1. Demographic and Clinical Characteristics of Training and Test Dataset

Variable*	**Training Dataset (n = 4,104)**	**Test Dataset (n = 1,026)**	**P-value****
Femoral Neck BMD	0.78 ± 0.13	0.79 ± 0.13	0.609
Total Hip BMD	0.96 ± 0.14	0.96 ± 0.14	0.55
Total Spine BMD	1.07 ± 0.19	1.07 ± 0.19	0.356
Age	73.77 ± 5.93	73.95 ± 5.83	0.784
Height	174.15 ± 6.81	174.16 ± 6.61	0.693
Weight	83.28 ± 13.46	82.73 ± 12.73	0.366
Alcohol use	4.25 ± 6.87	3.98 ± 6.06	0.3
Generic Risk Score	31.6 ± 0.44	31.59 ± 0.45	0.803
Impairment of Instrumental Activities of Daily Living	0.37 ± 0.88	0.36 ± 0.83	0.07
Walking Speed	1.07 ± 0.27	1.08 ± 0.29	0.53
Smoking (Current Smoking)	2.63%	3.51%	0.225
Race (White)	89.8%	90%	0.955

* Continuous variables were expressed as mean $\pm$ SD, and categorical variables were expressed as number (%).

** P-values were obtained by t – test for continuous variables and chi-square test for categorical variable.

Table 2. Comparisons of various ML models: Mean Square Error (MSE) and Mean Absolute Error (MAE) at each BMD with ***phenotype covariates + GRS*** in the test dataset (n = 1,026)

Femoral Neck BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0146	0.0126	0.0941	0.0881
Random Forest	0.0152	0.0151	0.0945	0.0960
Gradient Boosting	0.0142	0.0141	0.0946	0.0360
Neural Network	0.0150	0.0122	0.0958	0.0868

Total Hip BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0164	0.0151	0.1008	0.0967
Random Forest	0.0163	0.0163	0.1003	0.1005
Gradient Boosting	0.0162	0.0157	0.1026	0.0394
Neural Network	0.0162	0.0153	0.1005	0.0974

Total Spine BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0365	0.0316	0.1489	0.1363
Random Forest	0.0353	0.0353	0.1469	0.1439
Gradient Boosting	0.0349	0.0347	0.1490	0.0562
Neural Network	0.0352	0.0324	0.1460	0.1378

Table 3. Comparisons of various ML models: Mean Square Error (MSE) and Mean Absolute Error (MAE) at each BMD with ***phenotype covariates + 1,103 SNPs*** in the test dataset (n = 1,026)

Femoral Neck BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0146	0.0126	0.0941	0.0881
Random Forest	0.0152	0.0151	0.0945	0.0960
Gradient Boosting	0.0142	0.0141	0.0946	0.0360
Neural Network	0.0150	0.0122	0.0958	0.0868

Total Hip BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0185	0.0186	0.1079	0.1067
Random Forest	0.0180	0.0183	0.1061	0.1066
Gradient Boosting	0.0164	0.0172	0.1000	0.0998
Neural Network	0.0155	0.0162	0.0987	0.1002

Total Spine BMD	**MSE**		**MAE**	
Model	**Test**	**Train**	**Test**	**Train**
Linear Regression	0.0455	0.0470	0.1671	0.1675
Random Forest	0.0337	0.0341	0.1467	0.1442
Gradient Boosting	0.0328	0.0330	0.1444	0.1417
Neural Network	0.0378	0.0381	0.1565	0.1549

Table 4. Results of the Wilcoxon Signed-Rank Test for squared errors comprised by MSE between various ML models at each BMD with ***phenotype covariates + GRS*** in the test dataset (n = 1,026)

Femoral Neck BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	> .05	> .05	> .05
Gradient Boosting	> .05	< .0001	--
Random Forest	> .05	--	--

Total Hip BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	> .05	> .05	> .05
Gradient Boosting	> .05	> .05	--
Random Forest	> .05	--	--

Total Spine BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	> .05	> .05	> .05
Gradient Boosting	> .05	< .01	--
Random Forest	> .05	--	--

Table 5. Results of the Wilcoxon Signed-Rank Test for squared errors comprised by MSE between various ML models at each BMD with ***phenotype covariates + 1,103 SNPs*** in the test dataset (n = 1,026)

Femoral Neck BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	< .0001	> .05	< .0001
Gradient Boosting	< .0001	< .0001	--
Random Forest	< .0001	--	--

Total Hip BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	< .0001	< .0001	< .0001
Gradient Boosting	< .0001	> .05	--
Random Forest	< .0001	--	--

Total Spine BMD		***p* Value**	
Model	**Linear Regression**	**Random Forest**	**Gradient Boosting**
Neural Network	< .0001	< .0001	< .0001
Gradient Boosting	< .0001	< .001	--
Random Forest	< .0001	--	--

Legends

Figure 1. Overview of Data Process Flow to predict BMD.

Figure 2. Comparisons of various ML models: Mean Square Error vs. Number of iterations in the test dataset ($n = 1{,}026$). First row (A, B, C) shows the performance of each model with the predictors, *phenotype covariates + GRS*, in the test dataset at different BMD site. Second row (D, E, F) shows the performance of each model with the predictors, *phenotype covariates + 1,103 SNPs*, in the test dataset at different BMD site.

Figure 1.

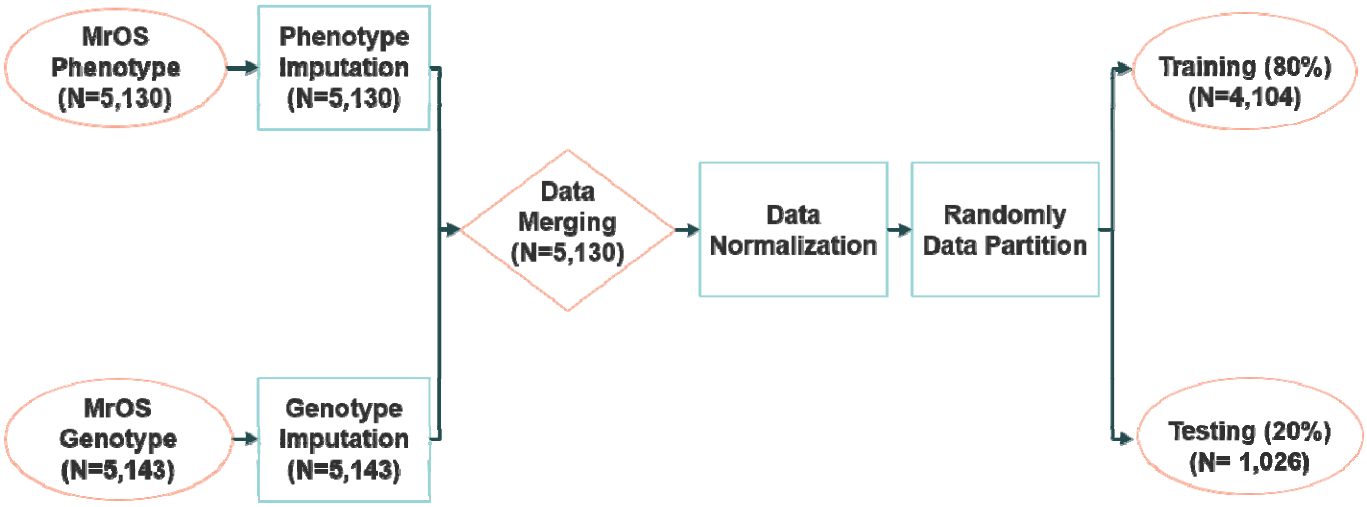
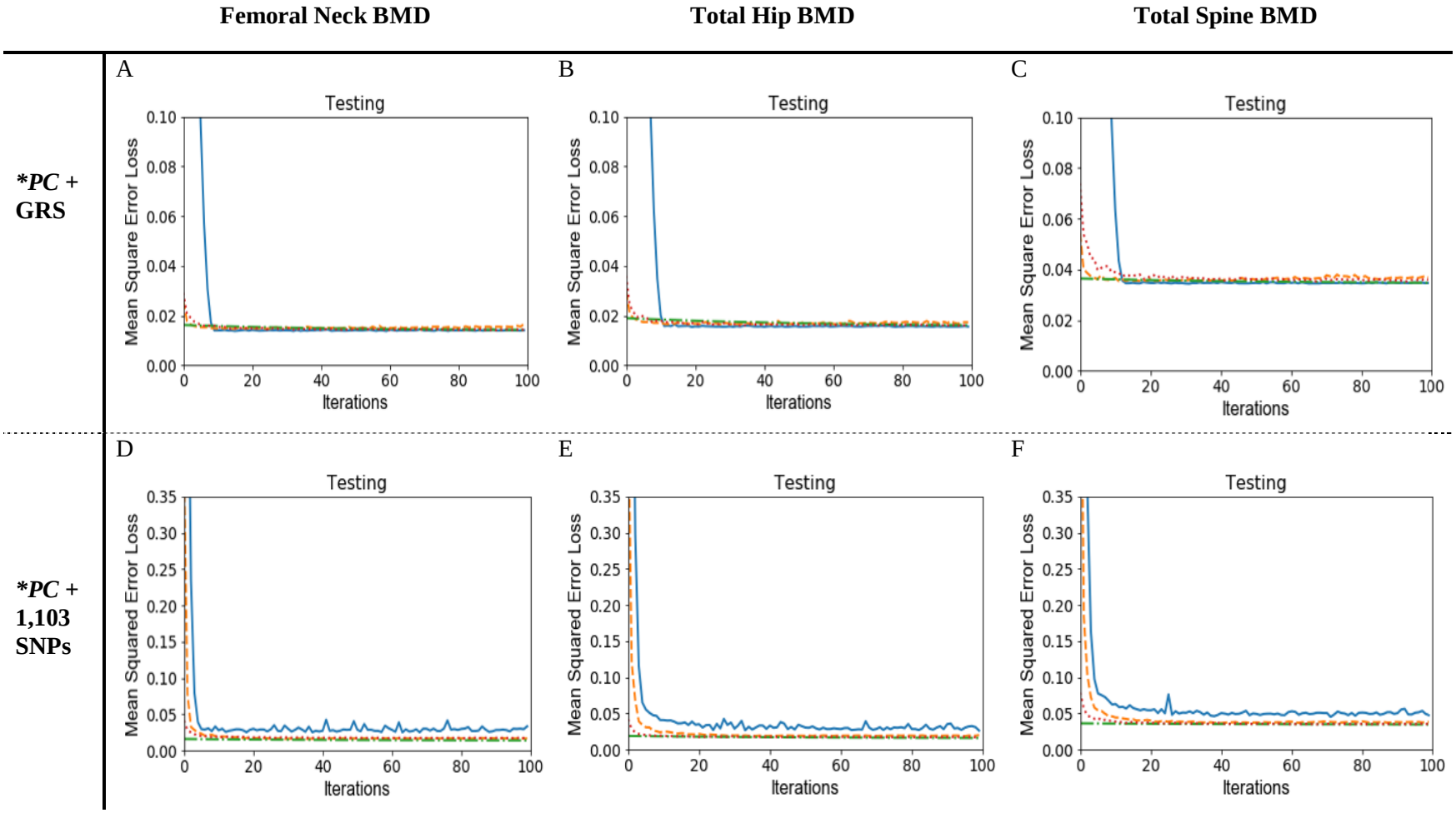

MrOS Phenotype (N=5,130)
Phenotype Imputation (N=5,130)
MrOS Genotype (N=5,143)
Genotype Imputation (N=5,143)
Data Merging (N=5,130)
Data Normalization
Randomly Data Partition
Training (80%) (N=4,104)
Testing (20%) (N= 1,026)

Figure 2.


Femoral Neck BMD
Total Hip BMD
Total Spine BMD
*PC + GRS
*PC + 1,103 SNPs
A
B
C
D
E
F
Testing
Mean Square Error Loss
Mean Squared Error Loss
Iterations

**PC*: The list of phenotype covariates are shown in Table 1