35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Osteoporosis Risk Prediction Using Machine Learning and Conventional Methods

Sung Kean Kim, Tae Keun Yoo, Ein Oh, Deok Won Kim, *Life member, IEEE**

***Abstract*— A number of clinical decision tools for osteoporosis risk assessment have been developed to select postmenopausal women for the measurement of bone mineral density. We developed and validated machine learning models with the aim of more accurately identifying the risk of osteoporosis in postmenopausal women, and compared with the ability of a conventional clinical decision tool, osteoporosis self-assessment tool (OST). We collected medical records from Korean postmenopausal women based on the Korea National Health and Nutrition Surveys (KNHANES V-1). The training data set was used to construct models based on popular machine learning algorithms such as support vector machines (SVM), random forests (RF), artificial neural networks (ANN), and logistic regression (LR) based on various predictors associated with low bone density. The learning models were compared with OST. SVM had significantly better area under the curve (AUC) of the receiver operating characteristic (ROC) than ANN, LR, and OST. Validation on the test set showed that SVM predicted osteoporosis risk with an AUC of 0.827, accuracy of 76.7%, sensitivity of 77.8%, and specificity of 76.0%. We were the first to perform comparisons of the performance of osteoporosis prediction between the machine learning and conventional methods using population-based epidemiological data. The machine learning methods may be effective tools for identifying postmenopausal women at high risk for osteoporosis.**

I. Introduction

Fracture due to osteoporosis is one of the major factors of disability and death in elderly people [1]. Osteoporosis is common in postmenopausal women but is asymptomatic until a fracture occurs. The World Health Organization (WHO) estimates that 30% of all postmenopausal women have osteoporosis, which is defined as bone mineral density (BMD) 2.5 standard deviations below the young healthy adult mean (T-score≤-2.5) [2]. Dual X-ray absorptionmetry (DEXA) of total hip, femoral neck, and lumbar spine is the most widely used tool for diagnosing osteoporosis. However, mass screening using DEXA is not widely recommended as it is a high-cost method of evaluating BMD [3]. Therefore, selecting patients for DEXA is an important task for cost-effective screening for osteoporosis.

A number of epidemiological studies have developed clinical decision tools for osteoporosis risk assessment to select postmenopausal women for the measurement of BMD. The purpose of these clinical decision tools is to help estimate the risk for osteoporosis, not to diagnose osteoporosis. The osteoporosis self-assessment tool (OST) is one of the clinical decision tools, which is a simple formula based on age and body weight [4]. Although OST uses only two factors to predict osteoporosis risk, it has been shown to have good sensitivity with an appropriate cutoff value [5]. However, the decision tool has the limitation of low accuracy for clinical use [6].

Machine learning is an area of artificial intelligence research which uses statistical methods for data classification. Several machine learning techniques have been applied in clinical settings to predict disease and have shown higher accuracy for diagnosis than classical methods [7]. Support vector machines (SVM), random forests (RF), and artificial neural networks (ANN) have been widely used approaches in machine learning [7].

The SVM is based on mapping data to a higher dimensional space through a kernel function and choosing the maximum-margin hyper-plane that separates training data [8]. RF grows many classification trees built from a random subset of predictors and bootstrap samples [9]. ANN is comprised of several layers and connections which mimic biological neural networks to construct complex classifiers [10]. Logistic regression (LR) is another machine learning technique. LR is the gold standard method for analyzing binary medical data because it provides not only a predictive result, but also yields additional information such as a diagnostic odds ratio [11].

In this study, we developed the prediction models for osteoporosis using various machine learning methods including SVM, RF, ANN, and LR. We compared the performance of machine learning methods and OST using accuracy and area under the curve (AUC) of the receiver operating characteristic (ROC).

II. Materials and methods

A. Data Source

We collected data from Korean postmenopausal women based on the Korea National Health and Nutrition Examination Surveys (KNHANES V-1) conducted in 2010. BMD was measured by DEXA using Hologic Discovery (Hologic Inc., Bedford, MA). Patients who were determined to have postmenopausal status were included in this study. We categorized the postmenopausal women into a control group and an osteoporotic group with low BMD (T-score≤-2.5) at

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A03045612).

S. K. Kim is with the Graduate Program in Biomedical Engineering, Yonsei University, Seoul, Korea (e-mail: sdm04sdm@yuhs.ac).

T. K. Yoo is with the Department of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: fawoo2@yuhs.ac).

E. Oh is with the Department of Medicine, Yonsei University College of Medicine, Seoul, Korea (e-mail: bluerose1186@gmail.com).

*D. W. Kim is a Professor with the Department of Medical Engineering, Yonsei University College of Medicine, Seoul, Korea (phone: 82-2-2228-1916; fax: 82-2-364-1572; e-mail: kdw@yuhs.ac).

any site among total hip, femoral neck, or lumbar spine measurements.

B. Data Analysis

The data were separated randomly into two independent data sets: training and test sets. The training set, comprised of 60% (1000 patients) of the entire dataset, was used to construct models based on SVM, RF, ANN, and LR. The clinical decision tool for screening osteoporosis, OST, was calculated according to its formula. The prediction models were internally validated using 10-fold cross validation [12]. We designed the 10-fold cross validation not only to assess performance, but also to optimize prediction models using machine learning techniques. We used 10-fold cross validation on the training set, and the performance was measured on the test set. The test set, comprised of 40% (674 patients) of the entire dataset, was used to assess ability to predict osteoporosis in postmenopausal women. Fig. 1 shows the overview of trained machine learning models to predict osteoporosis.

C. Model Selection and Validation

We used the 10-fold cross validation scheme to construct machine learning models. The purpose of the machine learning models was to predict osteoporosis risk using the health interview surveys concerning demographic characteristics and past histories listed in Table I. Due to high dimensionality, variable selection was a necessary technique to make an effective prediction model and to improve prediction performance [13]. We adopted a feature selection method of consistency subset evaluation for SVM, RF, and ANN [7], [14]. We determined the order of the variables with the embedded method of each machine learning method and decreased the number of variables to determine the best subset using backward elimination [13]. The remaining features that indicated the highest accuracy in 10-fold cross validation were the selected subset for prediction. For LR, we used the backward stepwise method for variable selection.

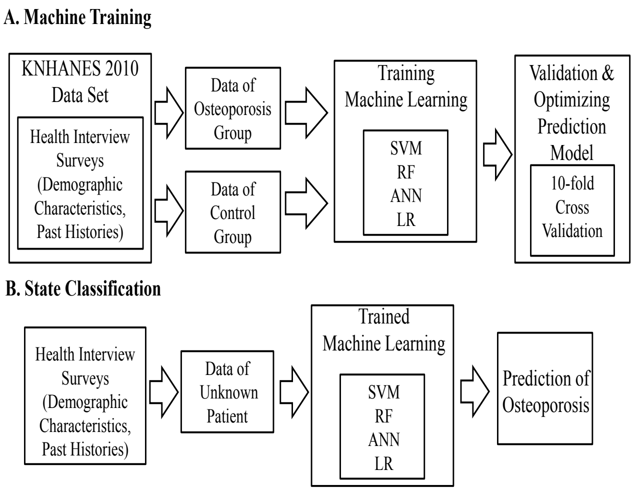
A. Machine Training
KNHANES 2010 Data Set
Health Interview Surveys (Demographic Characteristics, Past Histories)
Data of Osteoporosis Group
Data of Control Group
Training Machine Learning
SVM
RF
ANN
LR
Validation & Optimizing Prediction Model
10-fold Cross Validation
B. State Classification
Health Interview Surveys (Demographic Characteristics, Past Histories)
Data of Unknown Patient
Trained Machine Learning
SVM
RF
ANN
LR
Prediction of Osteoporosis

Figure 1. Overview of models to predict osteoporosis. The flow of training machine learning models (A) and the flow of state classification with unknown data (B).

TABLE I. DEMOGRAPHIC AND CLINICAL CHARACTERISTICS OF ANALYZED POSTMENOPAUSAL WOMEN

Variable*	**Without osteoporosis (n = 1091)**	**With osteoporosis at any site (n = 583)**	***P-value*†**
Age (years)	59.9±8.4	69.7±9.0	< 0.001
Height (cm)	154.6±5.4	150.3±5.7	< 0.001
Weight (kg)	58.9±8.3	52.9±8.2	< 0.001
BMI (kg/m^2)	24.6±3.3	23.3±3.1	< 0.001
Waist circumference (cm)	82.8±9.2	80.8±9.2	< 0.001
Pregnancy	4.3±2.2	5.0±2.4	< 0.001
Duration of menopause (years)	11.2±8.8	21.5±10.6	< 0.001
Duration of breast feeding (months)	43.3±44.4	74.5±57.3	< 0.001
Estrogen therapy	224 (20.5)	48 (8.2)	< 0.001
Hypertension	410 (37.5)	257 (44.0)	0.009
Hyperlipidemia	169 (15.4)	58 (9.9)	0.001
Diabetes mellitus	124 (11.3)	58 (9.9)	0.375
Osteoarthritis	278 (25.4)	174 (29.8)	0.055
Rheumatoid arthritis	31 (2.8)	25 (4.2)	0.116
History of fracture	144 (13.2)	99 (16.9)	0.036

*Table values are given as mean ± standard deviation or number (%) unless otherwise indicated.
†*P*-values were obtained by t-test and chi-square test.
BMI: body mass index

Data sets in this study were class-imbalanced because the control group contained significantly more samples than the osteoporotic group. Therefore, it was important to improve prediction models for the imbalanced data. To obtain the optimal result, we adopted a grid search in which a range of parameter values were tested using 10-fold cross validation strategy. We found the best classification model and employed its parameters for prediction. The optimal model of SVM was found using a Gaussian kernel function with a penalty parameter C of 100 and scaling factor σ of 30. In RF, the optimal number of trees was 100, and the number of predictors for each node was 3. The optimal ANN was set with 3 nodes of a hidden layer and learning rate of 0.1.

Due to the imbalanced data problem, prediction accuracy might not be a good criterion for assessing performance since the minor class has less influence on accuracy than the major class [15]. Therefore, we evaluated diagnostic abilities including not only accuracy, but also AUC, sensitivity, and specificity.

$$Sensitivity = TP / (TP + FN) \tag{1}$$

$$Specificity = TN / (TN + FP) \tag{2}$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{3}$$

True positive (TP): No. of osteoporotic women correctly identified as osteoporosis.

True negative (TN): No. of healthy women correctly identified as normal

False positive (FP): No. of healthy women incorrectly identified as osteoporosis

False negative (FN): No. of osteoporotic women incorrectly identified as normal

The AUC is known as a strong predictor of performance, especially with regard to imbalanced problems [16]. To compare the performance of models, we generated the ROC curves and selected cut-off points as the points on the ROC curve closest to the upper left corner. We used MATLAB 2010a (Inc., Natick, MA) for the analysis of machine learning and SPSS 18.0 (SPSS Inc., Chicago, IL) for LR and statistical analysis.

III. RESULTS

Five hundred eighty-three (34.8%) of 1674 postmenopausal women had combined osteoporosis at any site including total hip, femoral neck, or lumbar spine. Table I shows the characteristics of postmenopausal women categorized by the presence of osteoporosis. Table II summarizes the results of variable selection used in machine learning methods. While OST selected two variables to obtain simplicity, the machine learning methods except LR selected more than 10 variables for better performance. In 10-fold cross validation, we found that more complex discriminating functions such as SVM and RF showed better performance than simple linear functions such as LR and OST. For the AUCs, the SVM performed better than ANN (*p*=0.028), LR (*p*=0.037), and OST (*p*=0.037) using a Wilcoxon signed rank test.

Additionally, to assess the ability of the models for predicting osteoporosis, we applied our methods to a test set composed of the independent data. Table III shows the results of classifying the test set for selecting women at risk of osteoporosis. As a result, the SVM model was the best discriminator between controls and women with osteoporosis.

TABLE II. VARIABLE SELECTION IN VARIOUS MACHINE LEARNING METHODS FOR OSTEOPOROSIS RISK OF TOTAL HIP, FEMORAL NECK, OR LUMBAR SPINE

Variable	**Machine learning method**			
	SVM	**RF**	**ANN**	**LR**
Age	O	O	O	O
Height	O	O	O	
Weight	O	O	O	O
Body mass index	O	O	O	
Waist circumstance		O		
Pregnancy		O	O	
Duration of menopause	O	O	O	O
Duration of breast feeding	O	O	O	
Estrogen therapy	O			
Hypertension	O	O		
Hyperlipidemia	O	O		O
Diabetes mellitus	O	O	O	O
Osteoarthritis	O	O	O	O
Rheumatoid arthritis				
History of fracture			O	

SVM: support vector machines, RF: random forests, ANN: artificial neural networks, LR: logistic regression

TABLE III. DIAGNOSTIC PERFORMANCE OF OSTEOPOROSIS RISK ASSESSMENT FOR VARIOUS MACHINE LEARNING AND CONVENTIONAL CLINICAL METHODS

	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM	0.827	76.7	77.8	76.0
RF	0.824	76.5	76.6	76.5
ANN	0.807	75.2	76.6	74.4
LR	0.809	74.5	77.8	72.7
OST	0.806	74.0	75.4	73.2

AUC: area under the curve, SVM: support vector machines, RF: random forests, ANN: artificial neural networks, LR: logistic regression, OST: osteoporosis self-assessment tool

SVM predicted osteoporosis risk with an AUC of 0.827, an accuracy of 76.7%, sensitivity of 77.8%, and specificity of 76.0%. Fig. 2 shows the ROC curves of SVM, LR, and OST in predicting osteoporosis at any site. Because SVM had the highest AUC among the machine learning methods, we compared their ROC curves. LR was also included for comparison with SVM and OST. The AUCs of SVM, LR and OST were 0.827, 0.809, and 0.806, respectively (Table III).

IV. DISCUSSION AND CONCLUSION

We investigated a new approach based on machine learning techniques for predicting osteoporosis risk in postmenopausal women using data from the KNHANES V-1. We were the first to perform comparisons of the performance of osteoporosis prediction between the machine learning and conventional methods using population-based epidemiological data. Among the machine learning and conventional methods, our SVM model discriminated more accurately between women with osteoporosis and control women. In other words, SVM was more effective in analyzing the epidemiological underlying patterns of osteoporosis compared with the other methods.

Our proposed SVM model included age, height, weight, body mass index, duration of menopause, duration of breast feeding, estrogen therapy, hypertension, hyperlipidemia, diabetes mellitus, and osteoarthritis as predictors (Table II).

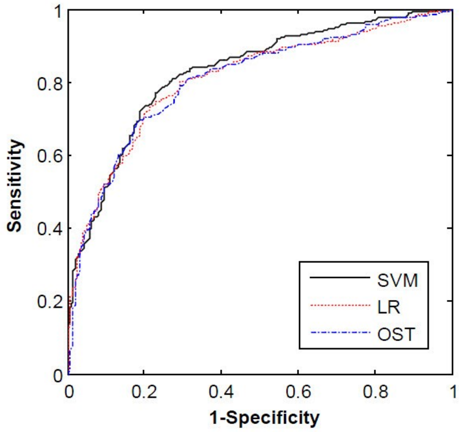

Figure 2. Receiver operating characteristic (ROC) curves of support vector machines (SVM), logistic regression (LR), and osteoporosis self-assessment

tool (OST) in predicting osteoporosis risk at any site among total hip, femoral neck, or lumbar spine

Similar to earlier studies concerning prediction for osteoporosis [4], [17], our results suggest that age and weight are most closely associated with the development of osteoporosis. However, our findings also demonstrated different factors involved in osteoporosis such as height, duration of menopause, duration of breast feeding, and presence of chronic diseases such as hypertension, hyperlipidemia, diabetes mellitus, and osteoarthritis. Our prediction model was able to consider these chronic diseases in combination using a SVM model characterized by nonlinearity and high dimension. Because the SVM model delicately handled a separating space composed of these factors in high dimension, it was possible to consider all factors for the improvement of sensitivity and specificity in predicting osteoporosis.

Women experience menopause at 50 years old on average [18]. Accordingly, when we regard the Korean women who are over 50 years old as potential menopausal population, menopausal women account for 31.8% of all women in Korea. The 31.8% corresponds to around 8.5 million [19]. Although our SVM showed small improvement of 2.7% in accuracy compared to OST, the 2.7% corresponds to approximately 230,000, which is not small population.

In conclusion, the most important finding of this study is the identification of postmenopausal women at high risk of osteoporosis to increase the possibility of appropriate treatment before fracture occurs. Machine learning methods might contribute to the advancement of clinical decision tools and understanding about the risk factors for osteoporosis. Further studies should be targeted at constructing an extended prediction model for progressive osteoporosis through the collection of prospective data, and the simultaneous prediction of osteopenia and osteoporosis using multi-category classification.

References

[1] D. Marshall, O. Johnell, and H. Wedel, "Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures," *BMJ*, vol. 312, pp. 1254-1259, May 1996.

[2] J. A. Kanis, "Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: synopsis of a WHO report," *Osteoporos Int*, vol. 4, pp. 368-381, Nov 1994.

[3] S. Nayak, M. S. Roberts, and S. L. Greenspan, "Cost-effectiveness of different screening strategies for osteoporosis in postmenopausal women," *Ann Intern Med*, vol. 155, pp. 751-761, Dec 2011.

[4] L. K. Koh, W. B. Sedrine, T. P. Torralba, A. Kung, S. Fujiwara, S. P. Chan, Q. R. Huang, R. Rajatanavin, K. S. Tsai, H. M. Park, and J. Y. Reginster, "A simple tool to identify Asian women at increased risk of osteoporosis," *Osteoporos Int*, vol. 12, pp. 699-705, Sep 2001.

[5] F. Richy, M. Gourlay, P. D. Ross, S. S. Sen, L. Radican, F. D. Ceulaer, W. B. Sedrine, O. Ethgen, O. Bruyere, and J. Y. Reginster, "Validation and comparative evaluation of the osteoporosis self-assessment tool (OST) in a Caucasian population from Belgium," *QJM*, vol. 97, pp. 39-46, Jan 2004.

[6] L. G. Raisz, "Screening for osteoporosis," *N Engl J Med*, vol. 353, pp. 164-171, Jul 2005.

[7] C. H. Hsieh, R. H. Lu, N. H. Lee, W. T. Chiu, M. H. Hsu, and Y. C. Li, "Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks," *Surgery*, vol. 149, pp. 87-93, Jan 2011.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, pp. 273-297 Sep 1995.

[9] L. Breiman, "Random forests," *Mach Learn*, vol. 45, pp. 5-32, Oct 2001.

[10] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *J Biomed Inform*, vol. 35, pp. 352-359, Oct, 2002.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[12] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *IJCAI*, vol. 14, pp. 1137-1145, 1995.

[13] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-2517, Oct 2007.

[14] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif Intell*, vol. 151, pp. 155-176, Dec 2003.

[15] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recogn*, vol. 40, pp. 3358-3378, Dec 2007.

[16] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int T Comput Sci Eng*, vol. 30, pp. 25-36, 2006.

[17] S. Khosla and L. J. Melton 3rd, "Osteopenia," *N Engl J Med*, vol. 356, pp. 2293-2300, May 2007.

[18] E. J. Yeun, "A study on the health promoting lifestyle practices of middle-aged women in Korea," *J Korean Soc Health Educ & Promo*, vol. 17, pp. 41-59, Mar 2000.

[19] Korean National Statistical Office 2010, http://www.kosis.kr