creative commons

C O M M O N S D E E D

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연세대학교
YONSEI UNIVERSITY

Establishment of hip fracture prediction model using radiomics texture analysis of dual-energy x-ray absorptiometry images with machine learning application

Namki Hong

Department of Medicine

The Graduate School, Yonsei University

연세대학교
YONSEI UNIVERSITY

연세대학교
YONSEI UNIVERSITY

Establishment of hip fracture prediction model using radiomics texture analysis of dual-energy x-ray absorptiometry images with machine learning application

Directed by Professor Yumie Rhee

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy in Medical Science

Namki Hong

December 2021

This certifies that the Doctoral Dissertation of Namki Hong is approved.

Yumie Rhee

Thesis Supervisor : Yumie Rhee

문성환

Thesis Committee Member#1 : Seong-Hwan Moon

김덕윤

Thesis Committee Member#2 : Deog-Yoon Kim

김창오

Thesis Committee Member#3: Chang Oh Kim

김휘영

Thesis Committee Member#4: Hwiyoung Kim

The Graduate School
Yonsei University

December 2021

ACKNOWLEDGEMENTS

This page is exclusively designed to note my gratitude and respect for those who helped me to complete my thesis. I am deeply indebted to my mentor and role model, Professor Yumie Rhee, for her warm encouragement, help, guidance, and support throughout this study. My journey into bone and mineral metabolism as an endocrinologist would not have been possible without her enthusiasm and passion for this field. I would like to express my sincere gratitude to thesis committee members, Professor Seong-Hwan Moon, Deog-Yoon Kim, Chang Oh Kim, and Hwiyoung Kim who provided fundamental insights to improve this study. I am indebted to my colleagues, Heajeong Park, Narae Park, Eunyoung Park, Sookyung Han, and Dawon Song for supporting data gathering. I thank coordinating committee of Korean Urban Rural Elderly cohort, Chang Oh Kim, Hyeon Chang Kim, and Yumie Rhee, and research team members, Jongeun Kim and Young-Mi Yoon, for allowing my participation into the study and access to the database. My family deserves endless gratitude: my wife Hyelin Lee for being love of my life and her endless patience for lazy husband, my children Jina Hong and Jinseo Hong for loving workaholic daddy without any conditions, and my parents for their love and wisdom that guided me. Last but not least, I thank God for his abundant, abounding grace over my life.

Namki Hong

<TABLE OF CONTENTS>

ABSTRACT ································· 1-2

I. INTRODUCTION ································· 3-4
II. MATERIALS AND METHODS ································· 5-33
1. Study subjects ································· 5
A. Derivation set ································· 5-6
B. Clinical validation set ································· 7
2. Image processing ································· 7-8
3. Feature extraction and selection in the train set ································· 9-30
4. Bone radiomics score model development ································· 31-32
5. Statistical analysis ································· 33
III. RESULTS ································· 34-45
1. Characteristics of study subjects ································· 34-35
2. Selected texture features and performances of DXA hip radiomics models ································· 36-37
3. Correlations of the bone radiomics score with clinical and hip geometry parameters ································· 38-39
4. Hip fracture risk prediction using the bone radiomics score in clinical validation set ································· 39-45
IV. DISCUSSION ································· 46-50
V. CONCLUSION ································· 50

REFERENCES ································· 51-55
ABSTRACT (IN KOREAN) ································· 56-57
PUBLICATION LIST ································· 58

연세대학교
YONSEI UNIVERSITY

LIST OF FIGURES

Figure 1. Study flow ········ 5
Figure 2. Development process of machine learning-based bone radiomics score for hip fracture prediction ········ 8
Figure 3. Removal of highly correlated feature ········ 25
Figure 4. Relative feature importance of selected 14 texture features from recursive feature elimination via elastic net model ········ 25
Figure 5. Performance of models in internal test set ········ 32
Figure 6. Schematic process of feature extraction and modeling of bone radiomics score. ········ 32
Figure 7. Normalized mean difference of texture features selected in final model between individuals who sustained hip fracture during follow-up and who did not in derivation set and clinical validation set ········ 37
Figure 8. Scatterplot of bone mineral density versus bone radiomics score in clinical validation set ········ 38
Figure 9. Risk of hip fracture according to bone radiomics score in the clinical validation set. Solid lines with caps indicate 95% confidence interval. ········ 40
Figure 10. Visual examples of DXA scans and radiomics feature values of a women with high radiomics score who sustained hip fracture versus a women with low radiomics score who did not sustain hip fracture during follow-up in clinical validation set ········ 41

LIST OF TABLES

Table 1. Formulas and descriptions of radiomics texture features ········ 9
Table 2. Inter-rater texture feature reproducibility by region-of-interest segmentation ········ 27
Table 3. Characteristics of the study subjects ········ 35
Table 4. Correlation between bone radiomics score and hip geometry parameters in clinical validation set ········ 39
Table 5. Predictors of incident hip fracture in clinical validation set ········ 42
Table 6. Gradient of risk (hazard ratio per SD) for femoral neck aBMD and bone radiomics score in clinical validation set ········ 43
Table 7. Subgroup analysis in clinical validation set ········ 44

연세대학교
YONSEI UNIVERSITY

ABSTRACT

Establishment of hip fracture prediction model using radiomics texture analysis of dual-energy x-ray absorptiometry images with machine learning application

Namki Hong

Department of Medicine
The Graduate School, Yonsei University

Directed by Professor Yumie Rhee

Dual-energy X-ray absorptiometry (DXA)-based bone mineral density testing is standard to diagnose osteoporosis to detect individuals at high risk of fracture. A radiomics approach to extract quantifiable texture features from DXA hip images may improve hip fracture prediction without additional costs. Here, I investigated whether bone radiomics scores from DXA hip images could improve hip fracture prediction in a community-based cohort of older women. The derivation set (143 women who sustained hip fracture [mean age 73, time to fracture median 2.1 years] vs. 290 age-matched women [mean age 73] who did not sustain hip fracture during follow-up [median 5.5 years]) were split to train set (75%) and test set (25% hold -out set). Among various models using 14 selected features out of 300 texture features mined from DXA hip images in train set, random forest model was selected as best model to build bone radiomics score (range 0 to 100)

연세대학교
YONSEI UNIVERSITY

based on the performance in the test set. In a community-based cohort (2029 women, mean age 71) as clinical validation set, bone radiomics score was calculated using model fitted in train set. A total of 34 participants (1.7%) sustained hip fracture during median follow-up of 5.4 years (mean bone radiomics score 40±16 vs. 28±12 in non-fractured, $p<0.001$). A one-point bone radiomics score increment was associated with an 4% elevated risk of incident hip fracture (adjusted hazard ratio [aHR] 1.04, $p=0.001$) after adjustment for age, BMI, previous history of fracture, and femoral neck T-score, with improved model fit when added to covariates (likelihood ratio $\chi 2$ 10.74, $p=0.001$). The association between bone radiomics score with incident hip fracture remained robust (adjusted HR 1.06, $p<0.001$) after adjustment for FRAX hip fracture probability. Bone radiomics scores estimated from texture features of DXA hip images have the potential to improve hip fracture prediction.

Key words: DXA, fracture risk assessment, osteoporosis

연세대학교
YONSEI UNIVERSITY

Establishment of hip fracture prediction model using radiomics texture analysis of dual-energy x-ray absorptiometry images with machine learning application

Namki Hong

Department of Medicine
The Graduate School, Yonsei University

Directed by Professor Yumie Rhee

I. INTRODUCTION

Hip fracture has become a significant health problem in the era of global aging.[1] Hip fracture is associated with increased mortality, morbidity, and economic burden.[2,3] Currently, areal bone mineral density (aBMD) testing via dual-energy X-ray absorptiometry (DXA) is the standard method for diagnosing osteoporosis; both clinical risk factors and aBMD are able to identify individuals at high risk for hip fracture.[4] However, half of fragility fractures occur in individuals without osteoporosis, which leaves room for improvement in fracture risk prediction.[5]

Radiomics refers to comprehensive, automated high-throughput mining of quantitative standard-of-care medical image features to capture disease characteristics that are difficult to identify by human vision alone; in turn, this supports clinical decision making with improved diagnostic and/or predictive performance.[6,7] Radiomics can quantify a large array of radiologic phenotypes, including textures and spatial heterogeneity of the bone.[6,8] Several studies suggest that substantial spatial heterogeneity is related to aging, exercise, and diseases of bone distribution and microarchitecture at the proximal femur, which may partly

연세대학교
YONSEI UNIVERSITY

contribute to the susceptibility to hip fracture in addition to bone mass alone.[9-12] A recent study showed spatiotemporal heterogeneity in pixel-wise BMD changes with aging using DXA hip images, which suggests the applicability of radiomics for DXA images.[13] If properly leveraged, radiomics can be useful to mine quantitative texture indices at the pixel-wise level that are related to hip fracture from DXA hip images; this has the potential to improve hip fracture risk prediction as add-on information to DXA aBMD in standard-of-care practice.

In this study, I aimed to investigate whether the bone radiomics score using texture features from DXA can improve hip fracture risk prediction in older women.

연세대학교
YONSEI UNIVERSITY

II. MATERIALS AND METHODS

1. Study subjects

Two separate data set (Figure 1) were analyzed in this study to develop machine learning-based models for the bone radiomics score (derivation set) and to test the clinical utility of the score for hip fracture prediction in a prospective community-based cohort dataset with the consideration of time-to-event information (clinical validation set). This study was approved by the Institutional Review Board (IRB) of Severance Hospital, Seoul, Korea (no. 4-2020-0884; 4-2012-0172). The study was conducted in compliance with the World Medical Association Declaration of Helsinki.

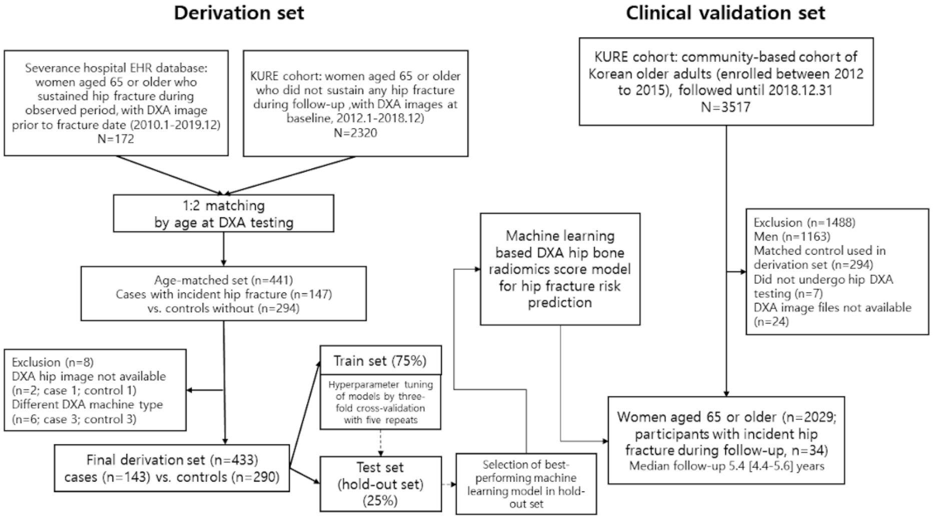

Derivation set
Clinical validation set
Severance hospital EHR database: women aged 65 or older who sustained hip fracture during observed period, with DXA image prior to fracture date (2010.1-2019.12) N=172
KURE cohort: women aged 65 or older who did not sustain any hip fracture during follow-up ,with DXA images at baseline, 2012.1-2018.12) N=2320
1:2 matching by age at DXA testing
Age-matched set (n=441) Cases with incident hip fracture (n=147) vs. controls without (n=294)
Exclusion (n=8) DXA hip image not available (n=2; case 1; control 1) Different DXA machine type (n=6; case 3; control 3)
Final derivation set (n=433) cases (n=143) vs. controls (n=290)
Train set (75%)
Hyperparameter tuning of models by three-fold cross-validation with five repeats
Test set (hold-out set) (25%)
Selection of best-performing machine learning model in hold-out set
Machine learning based DXA hip bone radiomics score model for hip fracture risk prediction
KURE cohort: community-based cohort of Korean older adults (enrolled between 2012 to 2015), followed until 2018.12.31 N=3517
Exclusion (n=1488) Men (n=1163) Matched control used in derivation set (n=294) Did not undergo hip DXA testing (n=7) DXA image files not available (n=24)
Women aged 65 or older (n=2029; participants with incident hip fracture during follow-up, n=34) Median follow-up 5.4 [4.4-5.6] years

Figure 1. Study flow. DXA, dual-energy X-ray absorptiometry; EHR, electronic health record; KURE, Korean Urban Rural Elderly.

A. Derivation set

To develop the bone radiomics score based on selected texture features which best discriminate individuals who sustain a hip fracture or not, electronic medical

연세대학교
YONSEI UNIVERSITY

records of women aged 65 years or older who sustained any fragility fracture at the hip between January 2010 and December 2019 with any DXA images prior to the fracture date (n=172) were retrieved from the Clinical Data Repository System of Severance Hospital, Seoul, Korea. All incident hip fracture events in derivation set were ascertained based on a medical record review conducted by investigators (NH and YR). The need for informed consent was waived by IRB due to the study design of retrospective medical record review. For control groups, women aged 65 years or older who did not sustain any hip fracture between January 2012 and December 2018 (n=2320) were selected as control candidates from a community-based prospective cohort of older Korean adults (Korean Urban Rural Elderly cohort, KURE). Written informed consent was obtained in all KURE cohort participants prior to the examination. DXA images were obtained at the time of enrollment (2012 to 2015). Hip fracture-free status during the follow-up period was determined by interviewer-assisted questionnaires during the second-wave follow-up with a 4-year interval (2016 to 2018). This was further ascertained by the absence of any International Classification of Diseases 10 codes (S72.0, S72.1) until December 31, 2018 using a database linked to the Korea National Claim Database of the Health Insurance Review and Assessment Service, covering 98% of entire Korean population. After 1:2 matching by age and follow-up duration, as well as exclusion of individuals without DXA hip images (n=2) or those with DXA images taken by an older version of the machine (n=6; QDR 4500, Hologic, USA), a total of 433 subjects (case n=143; control n=290) remained in the derivation set. The median time-to-fracture in the case group was 2.1 years with an interquartile range of 0.9 to 3.9. Follow-up duration (hip fracture-free duration) for control group was median 5.5 years [interquartile range 4.5-6.4] from the date of DXA testing until Dec 31, 2018. The derivation set was then split into a train set (75%) to fit models and an internal test set (a hold-out set, 25%) for evaluating model performance.

연세대학교
YONSEI UNIVERSITY

B. Clinical validation set

To test the performance of the bone radiomics score for hip fracture risk prediction in a community-based setting with the consideration of time-to-event information, the clinical validation set was constructed using the KURE cohort after excluding subset used as hip fracture-free controls to construct derivation set (Figure 1). Among a total of 3517 participants at baseline (2012 to 2015), men (n=1163), women who were used as matched control group in the derivation set (n=294), those who did not undergo hip DXA testing (n=7), and those without available DXA image files (n=24) were excluded. Data of remaining 2029 participants were analyzed as clinical validation set (median follow-up duration 5.4 years [interquartile range 4.4-5.6]; 34 participants sustained hip fracture during follow-up).

2. Image processing

DXA images were obtained according to the standardized protocol of the institution (Discovery W, Hologic, NH, USA). Because all investigation in KURE cohort study was conducted at the Severance Hospital, DXA images were obtained by a single, identical DXA machine (Discovery W, Hologic, NH, USA; fast array scan mode) in both derivation set and clinical validation set. Osteoporosis was defined as T-score -2.5 or lower at lumbar spine, femoral neck, and total hip in accordance with WHO criteria. Hip structure analysis was performed using Hologic APEX analysis software to derive hip geometry parameters. DXA hip reports of study subjects were retrieved in a DICOM file format from Severance Hospital Picture Archiving and Communication System (Figure 2).

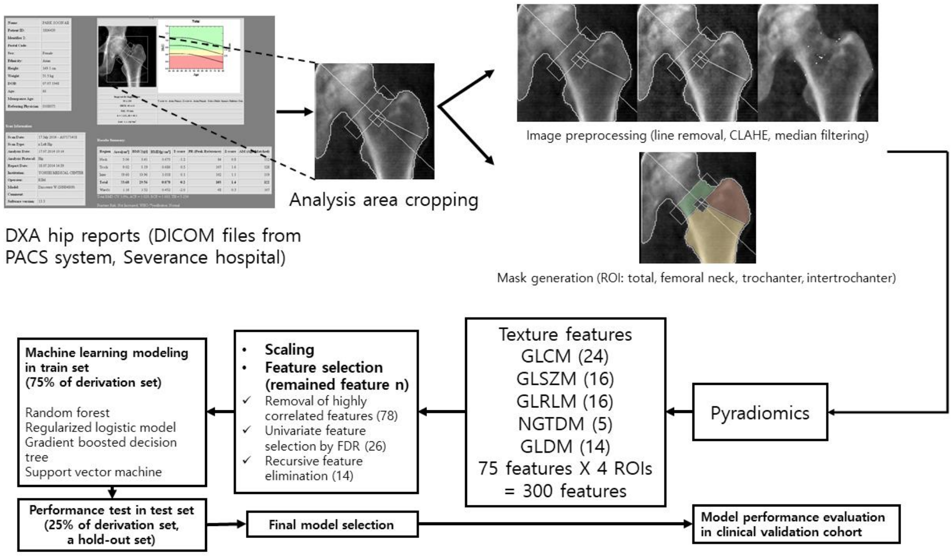

DXA hip reports (DICOM files from PACS system, Severance hospital)
Analysis area cropping
Image preprocessing (line removal, CLAHE, median filtering)
Mask generation (ROI: total, femoral neck, trochanter, intertrochanter)
Pyradiomics
Texture features
GLCM (24)
GLSZM (16)
GLRLM (16)
NGTDM (5)
GLDM (14)
75 features X 4 ROIs
= 300 features
Scaling
Feature selection (remained feature n)
Removal of highly correlated features (78)
Univariate feature selection by FDR (26)
Recursive feature elimination (14)
Machine learning modeling in train set (75% of derivation set)
Random forest
Regularized logistic model
Gradient boosted decision tree
Support vector machine
Performance test in test set (25% of derivation set, a hold-out set)
Final model selection
Model performance evaluation in clinical validation cohort

Figure 2. Development process of machine learning-based bone radiomics score for hip fracture prediction. Abbreviations: GLCM, gray level cooccurrence matrix; GLSZM, gray level size zone matrix; GLRLM, gray level run length matrix; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix.

The absence of artifacts or serious deviations in the region of interest in all DXA hip images were reviewed by two experts (NH, YR) with more than 10-years of practical experience. After cropping a box-shaped analysis area of the proximal femur (image size: mean 191 [95% CI 190-192] x mean 195 [95% CI 194-196] pixels in derivation set and mean 190 [95% CI 190-191] x mean 194 [95% CI 193-194] pixels in clinical validation set; pixel dimension: 0.504 mm), masks for regions-of-interest (ROIs: femoral neck, trochanter, intertrochanter, and total hip) were generated manually using open source software (3D Slicer, version 4.10.2, http://www.slicer.org)[14], as guided by existing lines in the images generated by the Hologic DXA machine. After automatic removal of the guidance lines penetrating ROIs by substituting line pixel values with median values of an adjacent 3-by-3 pixel region using our in-house python code, contrast-limited adaptive histogram equalization followed by median filtering was performed to enhance texture patterns in DXA hip images with reduction of background noise.[15,16]

3. Feature extraction and selection in the train set

For each ROI (femoral neck, trochanter, intertrochanter, total hip), 75 gray-level texture features were extracted via an automated process using Pyradiomics 3.0, an open-source python package (Table 1), yielding a total of 300 texture features per one DXA hip image.[17] Detailed formulas and descriptions of the texture features are provided in Table 1. The intraclass correlation (ICC) of texture features for inter-rater agreement (texture features from ROIs segmented by two independent analysts blinded to the outcome) was calculated from 30 randomly sampled DXA hip images, which showed robust reproducibility (95% confidence interval within an ICC range of 0.90 or higher) in 78% of 300 (233/300) texture features from four ROIs (Table 2).

Table 1. Formulas and descriptions of radiomics texture features[17]

Based methods	**Parameter**	**Formula**	**Description**
Grey Level Co-occurrence Matrix (GLCM) : A Gray Level Co-occurrence Matrix (GLCM) of size Ng×Ng describes the second-order joint probability function of an image region constrained by the mask and is defined as P(i,j|δ,θ). The (i,j)th element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image, that are separated by a distance of δ pixels along angle θ.			
GLCM Features	Autocorrelation	$autocorrelation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij$	Autocorrelation is a measure of the magnitude of the fineness and coarseness of texture.

연세대학교
YONSEI UNIVERSITY

Joint Average	$joint\ \ average = \mu_x = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)i$	The mean gray level intensity of the i distribution.
Cluster Prominence	$cluster\ \ prominence = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^4 p(i,j)$	Cluster Prominence is a measure of the skewness and asymmetry of the GLCM. A higher values implies more asymmetry about the mean while a lower value indicates a peak near the mean value and less variation about the mean.
Cluster Shade	$cluster\ \ shade = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^3 p(i,j)$	Cluster Shade is a measure of the skewness and uniformity of the GLCM. A higher cluster shade implies greater asymmetry about the mean.
Cluster Tendency	$cluster\ \ tendency = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j-\mu_x-\mu_y)^2 p(i,j)$	Cluster Tendency is a measure of groupings of voxels with similar gray-level values.
Contrast	$cluster\ \ tendency = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i+j)^2 p(i,j)$	Contrast is a measure of the local intensity variation, favoring values away from the diagonal. A larger value correlates with a greater disparity in intensity values among neighboring voxels.

연세대학교
YONSEI UNIVERSITY

Correlation	$correlation = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij - \mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}$	Correlation is a value between 0 (uncorrelated) and 1 (perfectly correlated) showing the linear dependency of gray level values to their respective voxels in the GLCM.
Difference Average	$difference\ \ average = \sum_{k=0}^{N_g-1} kp_{x-y}(k)$	Difference Average measures the relationship between occurrences of pairs with similar intensity values and occurrences of pairs with differing intensity values.
Difference Entropy	$difference\ \ entropy = \sum_{k=0}^{N_g-1} p_{x-y}(k)\log_2(p_{x-y}(k)+\epsilon)$	Difference Entropy is a measure of the randomness/variability in neighborhood intensity value differences.
Difference Variance	$difference\ \ variance = \sum_{k=0}^{N_g-1} (k-DA)^2 p_{x-y}(k)$	Difference Variance is a measure of heterogeneity that places higher weights on differing intensity level pairs that deviate more from the mean.
Joint Energy	$joint\ \ entropy = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p(i,j))^2$	Energy is a measure of homogeneous patterns in the image. A greater Energy implies that there are more instances of intensity value pairs in the image

		that neighbor each other at higher frequencies.
Joint Entropy	$joint\ entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2(p(i,j)+\epsilon)$	Joint entropy is a measure of the randomness/variability in neighborhood intensity values.
Informational Measure of Correlation (IMC) 1	$IMC1 = \frac{HXY - HXY1}{\max\{HX, HY\}}$	This class of features characterizes the textures of an image / object by creating a new matrix GLCM based on counting how often pairs of pixels with specific gray-level values and in a specified spatial relationship (distance and direction) occur in the image / object and then computing statistics from GLCM.IMC1 assesses the correlation between each and every probability distribution (quantifying the complexity of the texture). It represents the information of a single distribution.
Informational Measure of Correlation (IMC) 2	$IMC2 = \sqrt{1 - e^{-2(HXY2 - HXY)}}$	IMC2 also assesses the correlation between each and every probability distribution (quantifying the complexity of the texture). It

연세대학교
YONSEI UNIVERSITY

		represents the information of two of the distributions.
Inverse Difference Moment (IDM)	$IDM = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k^2}$	IDM is a measure of the local homogeneity of an image. IDM weights are the inverse of the Contrast weights.
Maximal Correlation Coefficient (MCC)	$MCC = \sqrt{\begin{matrix} second\ largest \\ eigenvalue\ of\ Q \end{matrix}}$ $Q(i,j) = \sum_{k=0}^{N_g} \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$	The Maximal Correlation Coefficient is a measure of complexity of the texture and 0≤MCC≤1.
Inverse Difference Moment Normalized (IDMN)	$IDMN = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k^2}{N_g^2}\right)}$	IDMN (inverse difference moment normalized) is a measure of the local homogeneity of an image. IDMN weights are the inverse of the Contrast weights.
Inverse Difference (ID)	$ID = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k}$	ID is another measure of the local homogeneity of an image. With more uniform gray levels, the denominator will remain low, resulting in a higher overall value.
Inverse Difference Normalized (IDN)	$IDN = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k}{N_g}\right)}$	IDN is another measure of the local homogeneity of an image
Inverse Variance	$inverse\ variance = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{k^2}$	Inverse of the variance calculated is taken
Maximum Probability	$maximum\ probability = \max(p(i,j))$	Maximum Probability is occurrences of the

연세대학교
YONSEI UNIVERSITY

			most predominant pair of neighboring intensity values.
	Sum Average	$sum\ \ average = \sum_{k=2}^{2N_g} p_{x-y}(k)k$	Sum Average measures the relationship between occurrences of pairs with lower intensity values and occurrences of pairs with higher intensity values.
	Sum Entropy	$sum\ \ entropy = \sum_{k=2}^{2N_g} p_{x-y}(k)\log_2(p_{x+y}(k)+\epsilon)$	Sum Entropy is a sum of neighborhood intensity value differences.
	Sum of Squares	$sum\ \ squares = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu_x)^2 p(i,j)$	Sum of Squares or Variance is a measure in the distribution of neigboring intensity level pairs about the mean intensity level in the GLCM.

Gray Level Size Zone Matrix (GLSZM): A Gray Level Size Zone (GLSZM) quantifies gray level zones in an image. A gray level zone is defined as a the number of connected voxels that share the same gray level intensity. A voxel is considered connected if the distance is 1 according to the infinity norm (26-connected region in a 3D, 8-connected region in 2D). In a gray level size zone matrix P(i,j) the (i,j)th element equals the number of zones with gray level i and size j appear in image.

GLSZM features	Small Area Emphasis (SAE)	$SAE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{P(i,j)}{j^2}}{N_z}$	SAE is a measure of the distribution of small size zones, with a greater value indicative of more smaller size zones and more fine textures.
	Large Area Emphasis (LAE)	$LAE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}P(i,j)j^2}{N_z}$	LAE is a measure of the distribution of large area size zones, with a greater value indicative of more larger size zones

연세대학교
YONSEI UNIVERSITY

		and more coarse textures.
Gray Level Non-Uniformity (GLN)	$GLN = \frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_s} P(i,j))^2}{N_z}$	GLN measures the variability of gray-level intensity values in the image, with a lower value indicating more homogeneity in intensity values.
Gray Level Non-Uniformity Normalized (GLNN)	$GLNN = \frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_s} P(i,j))^2}{N_z^2}$	GLNN measures the variability of gray-level intensity values in the image, with a lower value indicating a greater similarity in intensity values. This is the normalized version of the GLN formula.
Size-Zone Non-Uniformity (SZN)	$SZN = \frac{\sum_{j=1}^{N_s}(\sum_{i=1}^{N_g} P(i,j))^2}{N_z}$	SZN measures the variability of size zone volumes in the image, with a lower value indicating more homogeneity in size zone volumes.
Size-Zone Non-Uniformity Normalized (SZNN)	$SZNN = \frac{\sum_{j=1}^{N_s}(\sum_{i=1}^{N_g} P(i,j))^2}{N_z^2}$	SZNN measures the variability of size zone volumes throughout the image, with a lower value indicating more homogeneity among zone size volumes in the image. This is the normalized version of the SZN formula.
Zone Percentage	$ZP = \frac{N_z}{N_p}$	ZP measures the coarseness of the

연세대학교
YONSEI UNIVERSITY

(ZP)		texture by taking the ratio of number of zones and number of voxels in the ROI.
Gray Level Variance (GLV)	$GLN = \sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)(i-\mu)^2$	GLV measures the variance in gray level intensities for the zones.
Zone Variance (ZV)	$ZV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_s} p(i,j)(j-\mu)^2$	ZV measures the variance in zone size volumes for the zones.
Zone Entropy (ZE)	$ZE = \sum_{i=1}^{N_g}\sum_{j=1}^{N_S} p(i,j)\log_2(p(i,j)+\epsilon)$	ZE measures the uncertainty/randomness in the distribution of zone sizes and gray levels. A higher value indicates more heterogeneneity in the texture patterns.
Low Gray Level Zone Emphasis (LGLZE)	$LGLZE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$	LGLZE measures the distribution of lower gray-level size zones, with a higher value indicating a greater proportion of lower gray-level values and size zones in the image.
High Gray Level Zone Emphasis (HGLZE)	$HGLZE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\mathbf{P}(i,j)i^2}{N_z}$	HGLZE measures the distribution of the higher gray-level values, with a higher value indicating a greater proportion of higher gray-level values and size zones in the image.
Small Area Low Gray Level Emphasis	$SALGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{\mathbf{P}(i,j)}{i^2j^2}}{N_z}$	SALGLE measures the proportion in the image of the joint

연세대학교
YONSEI UNIVERSITY

	(SALGLE)		distribution of smaller size zones with lower gray-level values.
	Small Area High Gray Level Emphasis(SAHGLE)	$SAHGLE = \dfrac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\dfrac{\mathbf{P}(i,j)i^2}{j^2}}{N_z}$	SAHGLE measures the proportion in the image of the joint distribution of smaller size zones with higher gray-level values.
	Large Area Low Gray Level Emphasis (LALGLE)	$LALGLE = \dfrac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\dfrac{\mathbf{P}(i,j)j^2}{i^2}}{N_z}$	LALGLE measures the proportion in the image of the joint distribution of larger size zones with lower gray-level values.
	Large Area High Gray Level Emphasis (LAHGLE)	$LAHGLE = \dfrac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\mathbf{P}(i,j)i^2j^2}{N_z}$	LAHGLE measures the proportion in the image of the joint distribution of larger size zones with higher gray-level values.

Gray Level Run Length Matrix (GLRLM): A Gray Level Run Length Matrix (GLRLM) quantifies gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value. In a gray level run length matrix P(i,j|θ), the (i,j)th element describes the number of runs with gray level i and length j occur in the image (ROI) along angle θ.

GLRLM features	Short Run Emphasis (SRE)	$SRE = \dfrac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\dfrac{\mathbf{P}(i,j\vert\theta)}{j^2}}{N_r(\theta)}$	SRE is a measure of the distribution of short run lengths, with a greater value indicative of shorter run lengths and more fine textural textures.
	Long Run Emphasis (LRE)	$LRE = \dfrac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j\vert\theta)j^2}{N_r(\theta)}$	LRE is a measure of the distribution of long run lengths, with a greater value indicative of longer

연세대학교
YONSEI UNIVERSITY

		run lengths and more coarse structural textures.
Gray Level Non-Uniformity (GLN)	$GLN = \frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_r}\mathbf{P}(i,j\|\theta))^2}{N_r(\theta)}$	GLN measures the similarity of gray-level intensity values in the image, where a lower GLN value correlates with a greater similarity in intensity values.
Gray Level Non-Uniformity Normalized (GLNN)	$GLNN = \frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_r}\mathbf{P}(i,j\|\theta))^2}{N_r(\theta)^2}$	GLNN measures the similarity of gray-level intensity values in the image, where a lower GLNN value correlates with a greater similarity in intensity values. This is the normalized version of the GLN formula.
Run Length Non-Uniformity (RLN)	$RLN = \frac{\sum_{j=1}^{N_r}(\sum_{i=1}^{N_g}\mathbf{P}(i,j\|\theta))^2}{N_r(\theta)}$	RLN measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image.
Run Length Non-Uniformity Normalized (RLNN)	$RLNN = \frac{\sum_{j=1}^{N_r}(\sum_{i=1}^{N_g}\mathbf{P}(i,j\|\theta))^2}{N_r(\theta)^2}$	RLNN measures the similarity of run lengths throughout the image, with a lower value indicating more homogeneity among run lengths in the image. This is the normalized version of the RLN formula.

연세대학교
YONSEI UNIVERSITY

Run Percentage (RP)	$RLNN = \frac{N_r(\theta)}{N_p}$	RP measures the coarseness of the texture by taking the ratio of number of runs and number of voxels in the ROI.
Gray Level Variance (GLV)	$GLV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)(i-\mu)^2$	GLV measures the variance in gray level intensity for the runs.
Run Variance (RV)	$RV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)(j-\mu)^2$	RV is a measure of the variance in runs for the run lengths.
Run Entropy (RE)	$RE = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j\mid\theta)\log_2(p(i,j\mid\theta) + \epsilon)$	RE measures the uncertainty/randomness in the distribution of run lengths and gray levels. A higher value indicates more heterogeneity in the texture patterns.
Low Gray Level Run Emphasis (LGLRE)	$LGLRE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j\mid\theta)}{i^2}}{N_r(\theta)}$	LGLRE measures the distribution of low gray-level values, with a higher value indicating a greater concentration of low gray-level values in the image.
High Gray Level Run Emphasis (HGLRE)	$HGLRE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j\mid\theta)\, i^2}{N_r(\theta)}$	HGLRE measures the distribution of the higher gray-level values, with a higher value indicating a greater concentration of high gray-level values in the image.
Short Run Low Gray Level Emphasis	$SRLGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j\mid\theta)}{i^2 j^2}}{N_r(\theta)}$	SRLGLE measures the joint distribution of shorter run lengths

연세대학교
YONSEI UNIVERSITY

	(SRLGLE)		with lower gray-level values.
	Short Run High Gray Level Emphasis	$SBHGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j\mid\theta)i^2}{j^2}}{N_r(\theta)}$	SRHGLE measures the joint distribution of shorter run lengths with higher gray-level values.
	Long Run Low Gray Level Emphasis (LRLGLE)	$LRLGLRE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j\mid\theta)j^2}{i^2}}{N_r(\theta)}$	LRLGLRE measures the joint distribution of long run lengths with lower gray-level values.
	Long Run High Gray Level Emphasis (LRHGLE)	$LRHGLRE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j\mid\theta)\, i^2 j^2}{N_r(\theta)}$	LRHGLRE measures the joint distribution of long run lengths with higher gray-level values.

Gray Level Dependence Matrix (GLDM) : A Gray Level Dependence Matrix (GLDM) quantifies gray level dependencies in an image. A gray level dependency is defined as a the number of connected voxels within distance δ that are dependent on the center voxel. A neighboring voxel with gray level j is considered dependent on center voxel with gray level i if |i−j|≤α. In a gray level dependence matrix P(i,j) the (i,j)th element describes the number of times a voxel with gray level i with j dependent voxels in its neighborhood appears in image.

GLDM features	Small Dependence Emphasis (SDE)	$SDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$	A measure of the distribution of small dependencies, with a greater value indicative of smaller dependence and less homogeneous textures.
	Large Dependence Emphasis (LDE)	$LDE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\mathbf{P}(i,j)j^2}{N_z}$	A measure of the distribution of large dependencies, with a greater value indicative of larger dependence and more homogeneous textures.

연세대학교
YONSEI UNIVERSITY

Gray Level Non-Uniformity (GLN)	$GLN = \frac{\sum_{i=1}^{N_g}(\sum_{j=1}^{N_d}\mathbf{P}(i,j))^2}{N_z}$	Measures the similarity of gray-level intensity values in the image, where a lower GLN value correlates with a greater similarity in intensity values.
Dependence Non-Uniformity (DN)	$DN = \frac{\sum_{j=1}^{N_d}(\sum_{i=1}^{N_g}\mathbf{P}(i,j))^2}{N_z}$	Measures the similarity of dependence throughout the image, with a lower value indicating more homogeneity among dependencies in the image.
Dependence Non-Uniformity Normalized (DNN)	$DNN = \frac{\sum_{j=1}^{N_d}(\sum_{i=1}^{N_g}\mathbf{P}(i,j))^2}{N_z^2}$	Measures the similarity of dependence throughout the image, with a lower value indicating more homogeneity among dependencies in the image. This is the normalized version of the DLN formula.
Gray Level Variance (GLV)	$GLV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_d} p(i,j)(i-\mu)^2 ,$ $where\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_d} ip(i,j)$	Measures the variance in grey level in the image.
Dependence Variance (DV)	$DV = \sum_{i=1}^{N_g}\sum_{j=1}^{N_d} p(i,j)(j-\mu)^2 ,$	Measures the variance in dependence size in

연세대학교
YONSEI UNIVERSITY

	$where\mu = \sum_{i=1}^{N_g}\sum_{j=1}^{N_d} jp(i,j)$	the image.
Dependence Entropy (DE)	$Dependence\ \ Entropy = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j)\log_2(p(i,j)+\epsilon)$	The randomness of GLDM. Higher Dependence Entropy implies more complex texture
Low Gray Level Emphasis (LGLE)	$LGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2}}{N_z}$	Measures the distribution of low gray-level values, with a higher value indicating a greater concentration of low gray-level values in the image.
High Gray Level Emphasis (HGLE)	$HGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)i^2}{N_z}$	Measures the distribution of the higher gray-level values, with a higher value indicating a greater concentration of high gray-level values in the image.
Small Dependence Low Gray Level Emphasis (SDLGLE)	$SDLGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)}{i^2j^2}}{N_z}$	Measures the joint distribution of small dependence with lower gray-level values.
Small Dependence High Gray Level Emphasis (SDHGLE)	$SDHGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)i^2}{j^2}}{N_z}$	Measures the joint distribution of small dependence with higher gray-level values.
Large Dependence Low Gray Level Emphasis (LDLGLE)	$LDLGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\frac{\mathbf{P}(i,j)j^2}{i^2}}{N_z}$	Measures the joint distribution of large dependence with lower gray-level values.

연세대학교
YONSEI UNIVERSITY

	Large Dependence High Gray Level Emphasis (LDHGLE)	$LDHGLE = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_d}\mathbf{P}(i,j)i^2j^2}{N_z}$	Measures the joint distribution of large dependence with higher gray-level values.

Neighboring Gray Tone Difference Matrix (NGTDM): A Neighboring Gray Tone Difference Matrix quantifies the difference between a gray value and the average gray value of its neighbors within distance δ.

NGTDM features	Coarseness	$Coarseness = \frac{1}{\sum_{i=1}^{N_g} p_i s_i}$	Coarseness is a measure of average difference between the center voxel and its neighborhood and is an indication of the spatial rate of change. A higher value indicates a lower spatial change rate and a locally more uniform texture.
	Contrast	$Contrast = \left(\frac{1}{N_{g,p}(N_{g,p}-1)}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_i p_j (i-j)^2\right)\left(\frac{1}{N_{v,p}}\sum_{i=1}^{N_g} s_i\right),$ where $p_i \neq 0, p_j \neq 0$	Contrast is a measure of the spatial intensity change, but is also dependent on the overall gray level dynamic range. Contrast is high when both the dynamic range and the spatial change rate are high, i.e. an image with a large range of gray levels, with large changes between voxels and their neighborhood.
	Busyness	$Busyness = \frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\lvert ip_i - jp_j\rvert},$ where $p_i \neq 0, p_j \neq 0$	A measure of the change from a pixel to its neighbor. A high value for busyness indicates a 'busy'

연세대학교
YONSEI UNIVERSITY

		image, with rapid changes of intensity between pixels and its neighborhood.
Complexity	$Complexity = \frac{1}{N_{v,p}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \lvert i - j \rvert \frac{p_i s_i + p_j s_j}{p_i + p_j},$ where $p_i \neq 0, p_j \neq 0$	An image is considered complex when there are many primitive components in the image, i.e. the image is non-uniform and there are many rapid changes in gray level intensity.
Strength	$Strength = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p_i + p_j)(i - j)^2}{\sum_{i=1}^{N_g} s_i},$ where $p_i \neq 0, p_j \neq 0$	Strength is a measure of the primitives in an image. Its value is high when the primitives are easily defined and visible, i.e. an image with slow change in intensity but more large coarse differences in gray level intensities.

After normalization of the feature scale, removal of highly correlated features (correlation coefficient > 0.80; Figure 3), univariate feature selection based on the false discovery rate, and recursive feature elimination via an elastic net model was performed to reduce the number of features and to select features that are relevant for predicting the risk of hip fracture in the train set (75% of the derivation cohort) (14 features, Figure 3; inter-rater ICC ≥ 0.90 in all selected texture features; Table 2). The relative feature importance from the recursive feature elimination process for selected radiomics features is shown in Figure 4.

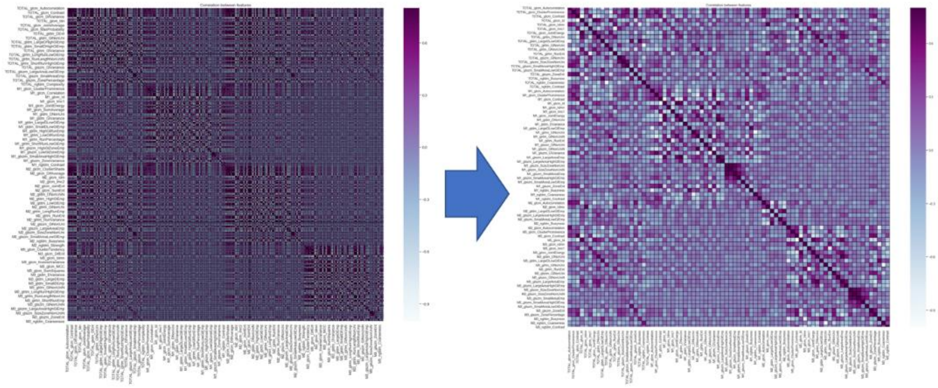


Figure 3. Removal of highly correlated feature. Right panel indicates correlation matrix after removal of highly correlated feature sets.

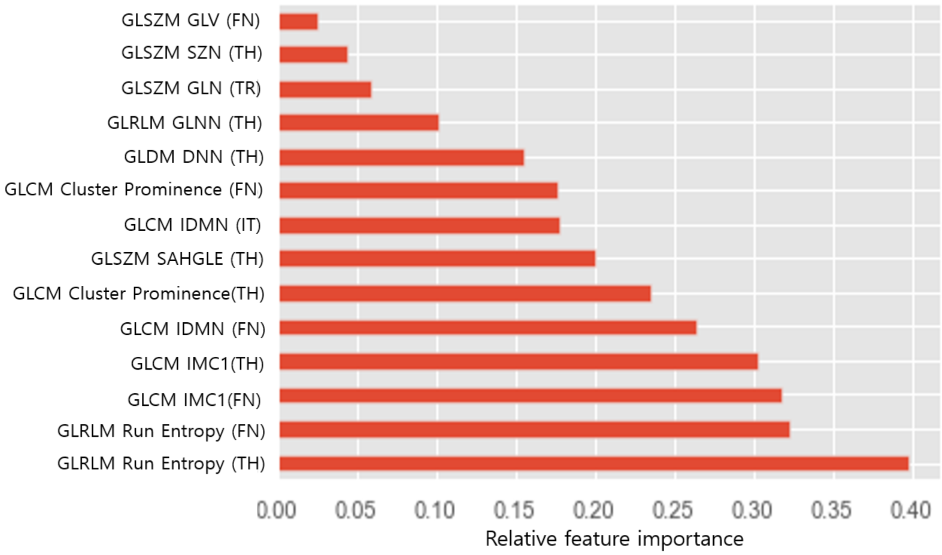


Figure 4. Relative feature importance of selected 14 texture features from recursive feature elimination via elastic net model (three-fold cross-validation

연세대학교
YONSEI UNIVERSITY

with five-time repetition in train set). Abbreviations: GLCM, gray level cooccurrence matrix; GLSZM, gray level size zone matrix; GLRLM, gray level run length matrix; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix; IDMN, inverse difference moment normalized; SAH GLE, small area high gray level emphasis; GLV, gray level variance; SZN, size zone uniformity; DNN, dependence non-uniformity normalized; GLN, gray level non-uniformity; GLNN, gray level non-uniformity normalized; IMC1, informational measure of correlation 1.

연세대학교
YONSEI UNIVERSITY

Table 2. Inter-rater texture feature reproducibility by region-of-interest segmentation

Texture features	Intraclass correlation (95% CI) Total hip	Femoral neck	Intertrochanter	Trochanter
Grey Level Co-occurrence Matrix (GLCM) features				
Autocorrelation	0.95 (0.90-0.97)	0.97 (0.94-0.98)	0.90 (0.79-0.95)	0.91 (0.81-0.95)
Joint Average	0.94 (0.88-0.97)	0.97 (0.93-0.98)	0.87 (0.73-0.93)	0.90 (0.79-0.95)
Cluster Prominence	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Cluster Shade	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Cluster Tendency	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Contrast	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Correlation	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Difference Average	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Difference Entropy	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Difference Variance	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Joint Energy	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Joint Entropy	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Informational Measure of Correlation (IMC) 1	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Informational Measure of Correlation (IMC) 2	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Inverse Difference Moment (IDM)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Maximal Correlation Coefficient (MCC)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Inverse Difference Moment Normalized (IDMN)	0.98 (0.95-0.99)	0.99 (0.98-0.99)	0.95 (0.90-0.98)	0.97 (0.94-0.99)
Inverse Difference (ID)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Inverse Difference Normalized (IDN)	0.98 (0.96-0.99)	0.98 (0.97-0.99)	0.96 (0.92-0.98)	0.98 (0.96-0.99)
Inverse Variance	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Maximum Probability	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Sum Average	0.94 (0.88-0.97)	0.97 (0.93-0.98)	0.87 (0.73-0.93)	0.90 (0.79-0.95)

Sum Entropy	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Sum of Squares	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Gray Level Size Zone Matrix (GLSZM) features				
Small Area Emphasis (SAE)	0.88 (0.76-0.94)	0.71 (0.40-0.86)	0.82 (0.63-0.91)	0.83 (0.65-0.92)
Large Area Emphasis (LAE)	0.99 (0.99-0.99)	0.98 (0.96-0.99)	0.99 (0.99-0.99)	0.98 (0.95-0.99)
Gray Level Non-Uniformity (GLN)	0.98 (0.95-0.99)	0.89 (0.78-0.95)	0.98 (0.96-0.99)	0.96 (0.93-0.98)
Gray Level Non-Uniformity Normalized (GLNN)	0.99 (0.98-0.99)	0.97 (0.95-0.98)	0.99 (0.98-0.99)	0.96 (0.93-0.98)
Size-Zone Non-Uniformity (SZN)	0.95 (0.90-0.97)	0.96 (0.92-0.98)	0.94 (0.87-0.97)	0.72 (0.40-0.86)
Size-Zone Non-Uniformity Normalized (SZNN)	0.92 (0.83-0.96)	0.91 (0.81-0.95)	0.88 (0.76-0.94)	0.62 (0.20-0.82)
Zone Percentage (ZP)	0.98 (0.97-0.99)	0.97 (0.93-0.98)	0.98 (0.97-0.99)	0.97 (0.93-0.98)
Gray Level Variance (GLV)	0.99 (0.99-0.99)	0.98 (0.96-0.99)	0.99 (0.98-0.99)	0.98 (0.97-0.99)
Zone Variance (ZV)	0.99 (0.99-0.99)	0.98 (0.96-0.99)	0.99 (0.99-0.99)	0.98 (0.96-0.99)
Zone Entropy (ZE)	0.97 (0.95-0.98)	0.96 (0.91-0.98)	0.98 (0.97-0.99)	0.95 (0.97-0.88)
Low Gray Level Zone Emphasis (LGLZE)	0.88 (0.76-0.94)	0.98 (0.95-0.99)	0.81 (0.61-0.91)	0.71 (0.40-0.86)
High Gray Level Zone Emphasis (HGLZE)	0.96 (0.92-0.98)	0.97 (0.95-0.98)	0.94 (0.88-0.97)	0.91 (0.81-0.95)
Small Area Low Gray Level Emphasis (SALGLE)	0.91 (0.82-0.96)	0.75 (0.48-0.88)	0.88 (0.74-0.94)	0.81 (0.61-0.91)
Small Area High Gray Level Emphasis (SAHGLE)	0.90 (0.76-0.94)	0.92 (0.84-0.96)	0.84 (0.67-0.92)	0.80 (0.58-0.90)
Large Area Low Gray Level Emphasis (LALGLE)	0.98 (0.97-0.99)	0.98 (0.97-0.99)	0.96 (0.93-0.98)	0.95 (0.91-0.97)
Large Area High Gray Level Emphasis (LAHGLE)	0.97 (0.94-0.98)	0.97 (0.95-0.98)	0.88 (0.75-0.94)	0.94 (0.87-0.97)

연세대학교
YONSEI UNIVERSITY

Gray Level Run Length Matrix (GLRLM) features				
Short Run Emphasis (SRE)	0.99 (0.98-0.99)	0.98 (0.96-0.99)	0.99 (0.98-0.99)	0.99 (0.98-0.99)
Long Run Emphasis (LRE)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Gray Level Non-Uniformity (GLN)	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.98-0.99)
Gray Level Non-Uniformity Normalized (GLNN)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Run Length Non-Uniformity (RLN)	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Run Length Non-Uniformity Normalized (RLNN)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Run Percentage (RP)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Gray Level Variance (GLV)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Run Variance (RV)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Run Entropy (RE)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Low Gray Level Run Emphasis (LGLRE)	0.92 (0.83-0.96)	0.93 (0.86-0.97)	0.74 (0.46-0.87)	0.81 (0.60-0.91)
High Gray Level Run Emphasis (HGLRE)	0.95 (0.90-0.97)	0.97 (0.94-0.98)	0.90 (0.80-0.95)	0.90 (0.80-0.95)
Short Run Low Gray Level Emphasis (SRLGLE)	0.90 (0.79-0.95)	0.94 (0.89-0.97)	0.75 (0.47-0.88)	0.72 (0.42-0.86)
Short Run High Gray Level Emphasis (SRHGLE)	0.97 (0.95-0.98)	0.98 (0.96-0.99)	0.95 (0.91-0.98)	0.94 (0.88-0.97)
Long Run Low Gray Level Emphasis (LRLGLE)	0.97 (0.95-0.99)	0.98 (0.97-0.99)	0.91 (0.82-0.95)	0.96 (0.92-0.98)
Long Run High Gray Level Emphasis (LRHGLE)	0.95 (0.90-0.97)	0.96 (0.92-0.98)	0.85 (0.70-0.93)	0.89 (0.78-0.95)
Gray Level Dependence Matrix (GLDM) features				
Small Dependence Emphasis (SDE)	0.99 (0.98-0.99)	0.97 (0.94-0.98)	0.99 (0.98-0.99)	0.97 (0.95-0.99)
Large Dependence Emphasis (LDE)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)

Gray Level Non-Uniformity (GLN)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Dependence Non-Uniformity (DN)	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Dependence Non-Uniformity Normalized (DNN)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Gray Level Variance (GLV)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Dependence Variance (DV)	0.99 (0.99-0.99)	0.97 (0.95-0.98)	0.99 (0.98-0.99)	0.98 (0.97-0.99)
Dependence Entropy (DE)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
Low Gray Level Emphasis (LGLE)	0.93 (0.86-0.96)	0.92 (0.84-0.96)	0.72 (0.41-0.86)	0.90 (0.79-0.95)
High Gray Level Emphasis (HGLE)	0.95 (0.90-0.97)	0.97 (0.94-0.98)	0.90 (0.79-0.95)	0.91 (0.81-0.95)
Small Dependence Low Gray Level Emphasis (SDLGLE)	0.91 (0.82-0.95)	0.90 (0.79-0.95)	0.70 (0.37-0.85)	0.83 (0.65-0.92)
Small Dependence High Gray Level Emphasis (SDHGLE)	0.96 (0.92-0.98)	0.98 (0.96-0.99)	0.94 (0.89-0.97)	0.91 (0.82-0.96)
Large Dependence Low Gray Level Emphasis (LDLGLE)	0.94 (0.88-0.97)	0.94 (0.89-0.97)	0.74 (0.46-0.87)	0.92 (0.84-0.96)
Large Dependence High Gray Level Emphasis (LDHGLE)	0.95 (0.89-0.97)	0.97 (0.94-0.98)	0.88 (0.74-0.94)	0.90 (0.80-0.95)
Neighboring Gray Tone Difference Matrix (NGTDM) features				
Coarseness	0.99 (0.99-0.99)	0.98 (0.97-0.99)	0.99 (0.99-0.99)	0.99 (0.98-0.99)
Contrast	0.98 (0.97-0.99)	0.97 (0.93-0.98)	0.97 (0.95-0.98)	0.98 (0.97-0.99)
Busyness	0.95 (0.90-0.97)	0.97 (0.95-0.98)	0.90 (0.80-0.95)	0.90 (0.79-0.95)
Complexity	0.99 (0.98-0.99)	0.97 (0.95-0.98)	0.97 (0.95-0.99)	0.98 (0.96-0.99)
Strength	0.99 (0.99-0.99)	0.99 (0.98-0.99)	0.99 (0.98-0.99)	0.99 (0.98-0.99)

4. Bone radiomics score model development

Models for hip fracture risk prediction based on 14 selected texture features were trained using four commonly used machine learning algorithms in the train set (75% of the derivation cohort): random forest, regularized logistic model (elastic net), gradient boosted decision tree, and support vector machine (scikit-learn, version 0.23.1).[18] Hyperparameters of each model were tuned using the grid search method with three-fold cross-validation for repeated five times in train set. In the internal test set (a hold-out set, 25% of the derivation cohort), model performance was evaluated using the area under the receiver-operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Model calibration was evaluated using the Brier score (mean squared difference between the observed event and predicted probability; ranged 0 to 1).[19,20] Lower Brier score (close to 0) indicates better calibration of the prediction. The bone radiomics score was calculated as the probability score ranged from 0 to 100 for hip fracture from the final model which showed the best performance (random forest) in the test set (Figure 5). Final model fitted in train set was applied to clinical validation set to calculate bone radiomics score (Figure 6).

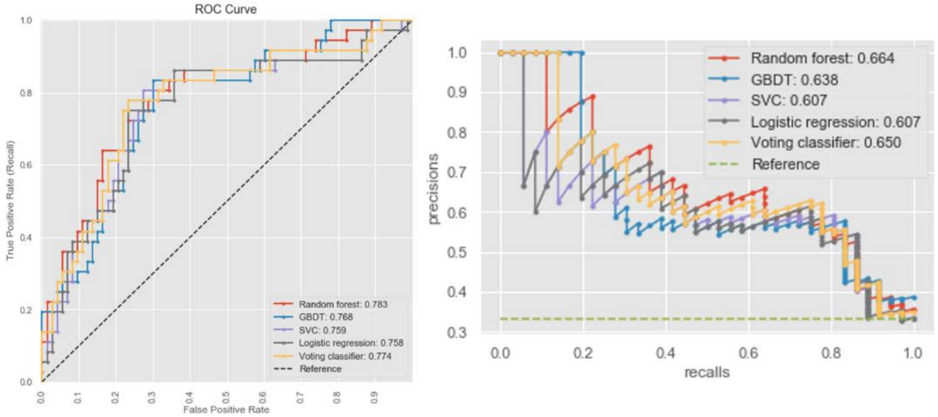


Figure 5. Performance of models in internal test set (a 25% hold-out set). Left panel: area under the receiver-operating characteristics curve (AUROC); right panel: area under the precision-recall curve (AUPRC).

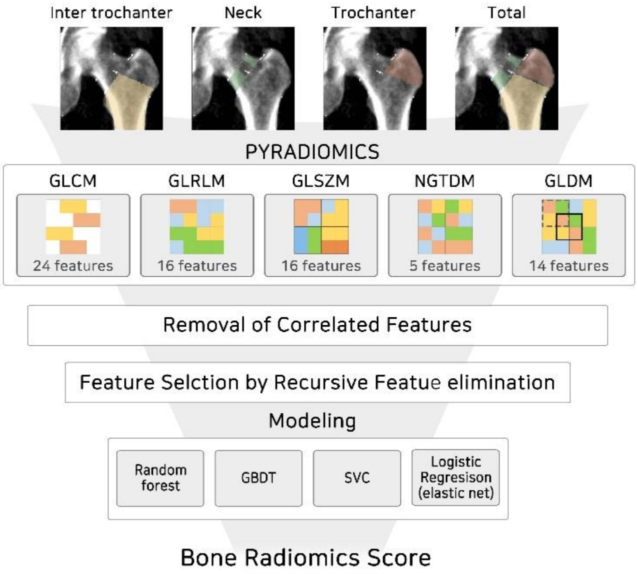


Figure 6. Schematic process of feature extraction and modeling of bone radiomics score

연세대학교
YONSEI UNIVERSITY

5. Statistical analysis

Group differences in continuous and categorical variables were compared using independent samples *t*-tests and Chi-squared tests, respectively. Risk of hip fracture was compared across groups of bone radiomics scores in the clinical validation set using the Cochran-Armitage test for trends.[21] None of the covariates violated the proportional hazard assumption. A Cox regression on bone radiomics model based on a sample size of 2029 subjects and event rate 1.7% observed in this study achieved 81% power at two-sided significance level 0.05 to detect a hazard ratio equal to 1.04.[22] Likelihood ratio test was used to compare the model fit between nested models with or without bone radiomics score. This study was written in accordance with the recommendation of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline.[23] All statistical analyses were performed using python and STATA 14.1 (College Station, TX, USA). The statistical significance level was set at two-sided 0.05.

III. RESULTS

1. Characteristics of the study subjects

In the derivation set, the mean age of subjects was 73 years (73 ± 6 in cases [n=143] vs. 73 ± 6 in controls [n=290], p=0.807). Median time to hip fracture in cases was 2.1 years [interquartile range 0.9 to 3.9] from the time point of DXA testing (Table 3). In controls who did not sustain hip fracture during follow-up, follow-up duration was 5.5 years [interquartile range 4.5-6.4] from DXA testing. Subjects who sustained hip fracture had a significantly lower femoral neck T-score (-2.8 ± 1.0 vs. -2.2 ± 1.0, p<0.001) and more history of fragility fracture (62% vs 33%, p<0.001) at baseline compared with controls. In the clinical validation set, subjects who sustained hip fracture (n=34) during follow-up were of an older age (75 ± 5 vs. 71 ± 4 years, p<0.001), had a higher prevalence of previous fracture (73 % vs. 33 %, p<0.001), osteoporosis (77% vs. 43%, p<0.001), and a lower femoral neck T-score (-2.8 ± 0.9 vs. -1.9 ± 0.9, p<0.001) compared to those who did not sustain any hip fracture (n=1995). The proportion of individuals with previous exposure to anti-osteoporosis medications within five years prior to DXA testing did not differ between individuals who experienced hip fracture and those who did not in both the derivation set and clinical validation set (p>0.05 for all).

연세대학교
YONSEI UNIVERSITY

Table 3. Characteristics of the study subjects

	Derivation set (Severance Hospital EHR database and community cohort)			Clinical validation set (Community-based cohort [KURE])		
	Case: subjects who sustained hip fracture during follow-up (n=143)	Control: subjects who did not sustain any hip fracture during follow-up (n=290)	P value	Sustained hip fracture during follow-up (n=34)	Without any hip fracture during follow-up (n=1995)	P value
Age	73 ± 6	73 ± 6	0.807	75 ± 5	71 ± 4	<0.001
BMI	23.2 ± 4.1	24.2 ± 3.1	0.004	23.6 ± 2.6	24.5 ± 3.1	0.111
Previous fracture	88 (62)	97 (33)	<0.001	25 (73)	33 (33)	<0.001
DXA T-score						
Lumbar spine	-2.1 ± 1.2	-2.0 ± 1.2	0.477	-1.9 ± 1.2	-1.8 ± 1.2	0.634
Femoral neck	-2.8 ± 1.0	-2.2 ± 1.0	<0.001	-2.8 ± 0.9	-1.9 ± 0.9	<0.001
Total hip	-2.0 ± 1.0	-1.3 ± 1.0	<0.001	-2.0 ± 0.8	-1.0 ± 0.9	<0.001
Osteoporosis	98 (69)	155 (53)	0.003	26 (77)	854 (43)	<0.001
Previous anti-osteoporosis medication use	31 (22)	67 (23)	0.739	10 (29)	477 (24)	0.456
Diabetes	46 (32)	54 (19)	0.002	8 (24)	456 (23)	0.926
Hypertension	93 (65)	187 (64)	0.910	26 (76)	1223 (61)	0.071
Any cancer	16 (11)	21 (7)	0.167	4 (11.8)	162 (8)	0.632

Data are presented as the mean ± standard deviation or number (%). Abbreviations: BMI, body mass index; DXA, dual-energy X-ray absorptiometry; EHR, electronic health record.

연세대학교
YONSEI UNIVERSITY

2. Selected texture features and performances of DXA hip radiomics models

In derivation set, the four models (random forest, gradient boosted decision tree, support vector classifier, and elastic net logistic regression) which were fitted using train set showed similar modest accuracy (0.72 to 0.74) and AUROCs (0.758 to 0.784) in the test set (hold-out set). Random forest model was chosen as the final model to estimate the bone radiomics score based on a numerically higher value in the AUROC (0.784; reference 0.500), AUPRC (0.664; reference 0.333), and F1-score (0.64). In clinical validation set, AUROC, AUPRC, and F1-score was 0.705 (reference 0.500), 0.094 (reference 0.017), and 0.12, respectively. Among the top 14 texture features used to develop the models, the GLRLM run entropy of the total hip, GLRLM run entropy of the femoral neck, GLCM informal measure of correlation 1 (IMC1) of the femoral neck and total hip, GLCM inverse difference moment normalized (IDMN) of the femoral neck, and GLCM cluster prominence of the total hip were the top five features with high importance contributing to the random forest model (Figure 3). Individuals who sustained a hip fracture had higher GLRLM run entropy, lower GLCM IMC1 (both indicate more heterogeneity in texture patterns), and higher GLCM IDMN with lower GLSZM GLN (both indicate less variation in gray-level pixel intensity values), compared to those who did not sustain a hip fracture in both the derivation set and clinical validation set (Figure 7).

연세대학교
YONSEI UNIVERSITY

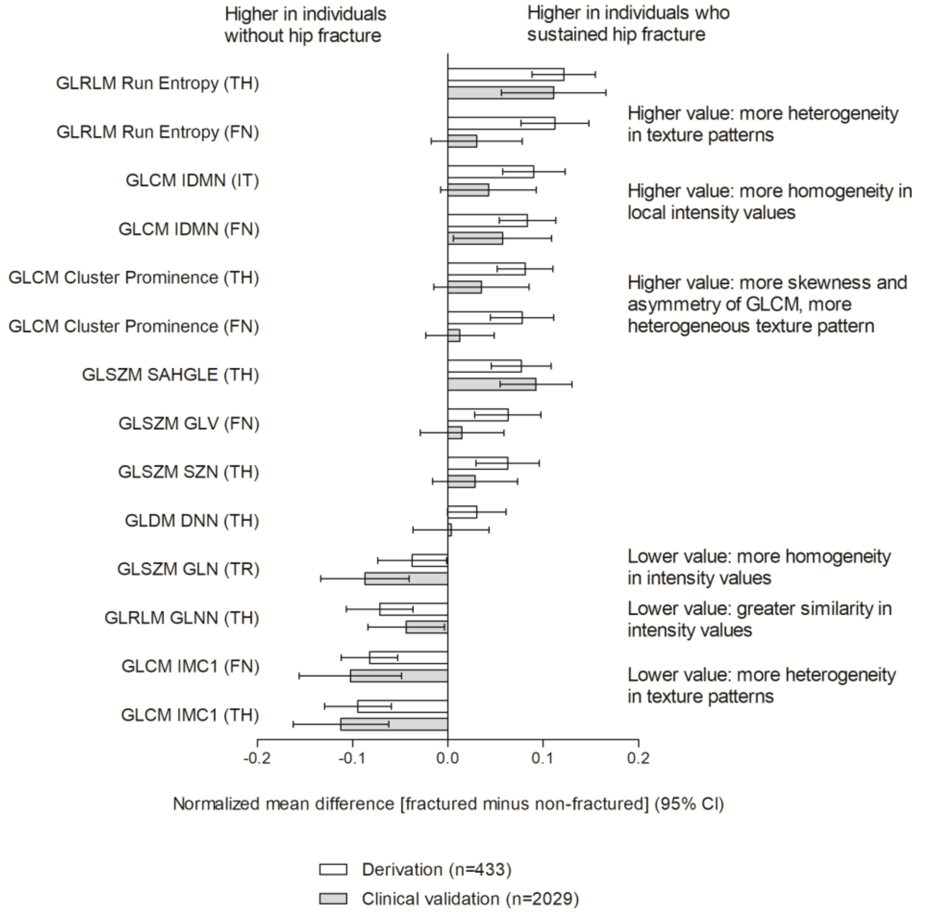


Figure 7. Normalized mean difference of texture features selected in final model between individuals who sustained hip fracture during follow-up and who did not in derivation set and clinical validation set. Abbreviations: GLCM, gray level cooccurrence matrix; GLSZM, gray level size zone matrix; GLRLM, gray level run length matrix; NGTDM, neighboring gray tone difference matrix; GLDM, gray level dependence matrix; IDMN, inverse difference moment normalized; SAH GLE, small area high gray level emphasis; GLV, gray level variance; SZN, size zone uniformity; DNN, dependence non-uniformity normalized; GLN, gray level non-uniformity; GLNN, gray level non-uniformity normalized; IMC1, informational measure of correlation 1.

3. Correlations of the bone radiomics score with clinical and hip geometry parameters

In the clinical validation set, the bone radiomics score ranged from 12 to 72, with mean score of 28 and standard deviation 12. The bone radiomics score showed a weak to moderate correlation with age (r=0.15, p<0.001), height (r=0.18, p<0.001), femoral neck BMD (r=-0.29, p<0.001), total hip BMD (r=-0.30, p<0.001; Figure 8), and FRAX hip fracture probabilities (r=0.22, p<0.001), whereas lumbar spine BMD (r=-0.08, p<0.001) and weight (r=0.01, p=0.747) showed negligible or no correlation with the bone radiomics score.

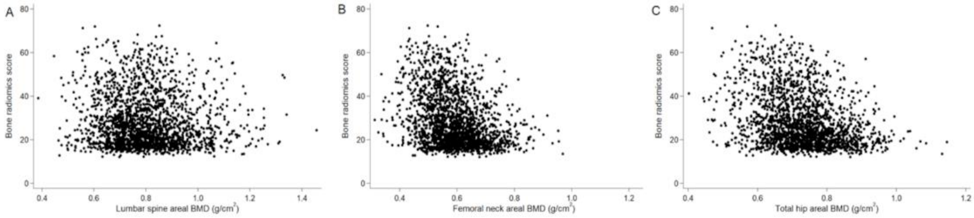


Figure 8. Scatterplot of bone mineral density versus bone radiomics score in clinical validation set

Higher bone radiomics scores were observed in individuals who had a prior history of fragility fracture (30 ± 13 vs. 27 ± 11, p<0.001) or osteoporosis (31 ± 13 vs. 26 ± 10, p<0.001) compared to those without. Bone radiomics score showed weak to moderate positive correlation with hip axis length (HAL, r=0.10), subperiosteal width (SPW, r=0.26 to 0.39), endocortical width (ECW, r=0.26 to 0.41), buckling ratio (BR, r=0.24 to 0.38), and negative correlation with cortical thickness (CT, r=-0.30 to -0.18) at narrow neck, intertrochanter, and femur shaft (Table 4).

연세대학교
YONSEI UNIVERSITY

Table 4. Correlation between bone radiomics score and hip geometry parameters in clinical validation set

Pearson correlation coefficient, r	SPW	ECW	CSA	CSMI	Z	CT	BR
Narrow neck	0.39*	0.41*	-0.12*	0.16*	-0.01	-0.30*	0.38*
Intertrochanter	0.26*	0.34*	-0.18*	0.05*	-0.06*	-0.19*	0.38*
Femur shaft	0.26*	0.26*	-0.05*	0.19*	0.01	-0.18*	0.24*
Hip axis length	0.10*						
Neck shaft angle	0.01						

Abbreviations: SPW, subperiosteal width; ECW, endocortical width; CSA, cross-sectional area; CSMI, cross-sectional momentum of inertia; Z, section modulus; CT, cortical thickness; BR, buckling ratio. *: p value< 0.05

4. Hip fracture risk prediction using the bone radiomics score in clinical validation set

In the clinical validation cohort, the risk of incident hip fracture during follow-up increased across bone radiomics score groups (p for the trend <0.001, Figure 9).

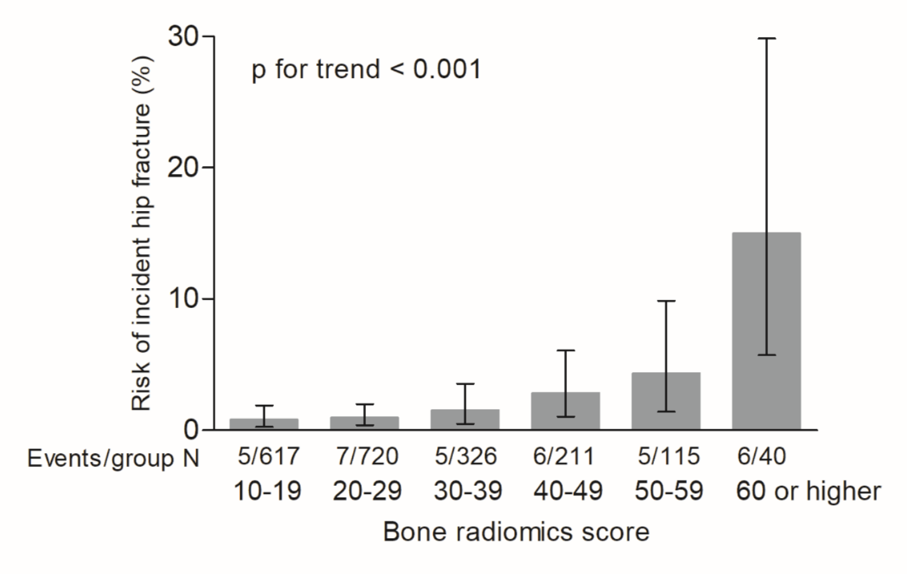
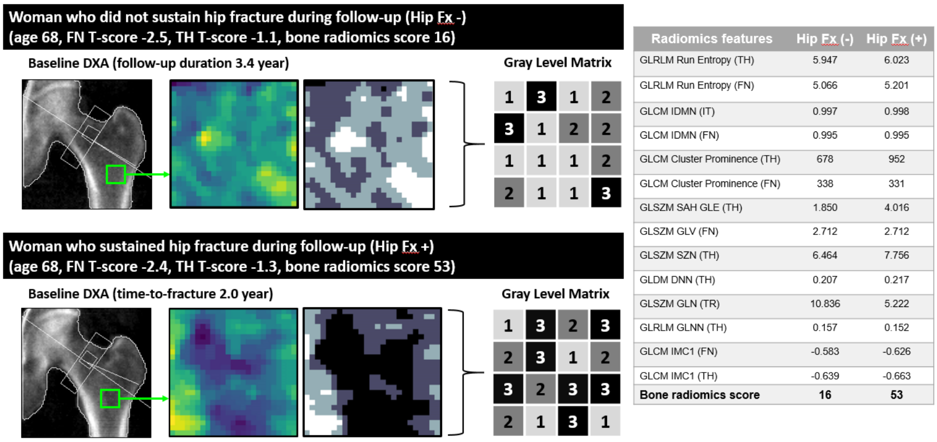


Figure 9. Risk of hip fracture according to bone radiomics score in the clinical validation set. Solid lines with caps indicate 95% confidence interval.

Individuals who sustained hip fracture during follow-up had higher bone radiomics score than those who did not (40 ± 16 vs. 28 ± 12, p<0.001). Figure 10 provides visual examples of DXA scans and radiomics feature values of participants with high radiomics score who sustained hip fracture and those with low radiomics score who did not sustain hip fracture during follow-up.



Woman who did not sustain hip fracture during follow-up (Hip Fx -)
(age 68, FN T-score -2.5, TH T-score -1.1, bone radiomics score 16)

Baseline DXA (follow-up duration 3.4 year)

Gray Level Matrix

1	3	1	2
3	1	2	2
1	1	1	2
2	1	1	3

Woman who sustained hip fracture during follow-up (Hip Fx +)
(age 68, FN T-score -2.4, TH T-score -1.3, bone radiomics score 53)

Baseline DXA (time-to-fracture 2.0 year)

Gray Level Matrix

1	3	2	3
2	3	1	2
3	2	3	3
2	1	3	1

Radiomics features	Hip Fx (-)	Hip Fx (+)
GLRLM Run Entropy (TH)	5.947	6.023
GLRLM Run Entropy (FN)	5.066	5.201
GLCM IDMN (IT)	0.997	0.998
GLCM IDMN (FN)	0.995	0.995
GLCM Cluster Prominence (TH)	678	952
GLCM Cluster Prominence (FN)	338	331
GLSZM SAH GLE (TH)	1.850	4.016
GLSZM GLV (FN)	2.712	2.712
GLSZM SZN (TH)	6.464	7.756
GLDM DNN (TH)	0.207	0.217
GLSZM GLN (TR)	10.836	5.222
GLRLM GLNN (TH)	0.157	0.152
GLCM IMC1 (FN)	-0.583	-0.626
GLCM IMC1 (TH)	-0.639	-0.663
Bone radiomics score	**16**	**53**

Figure 10. Visual examples of DXA scans and radiomics feature values of a woman with high radiomics score who sustained hip fracture versus a woman with low radiomics score who did not sustain hip fracture during follow-up in clinical validation set. Abbreviations: TH, total hip; FN, femoral neck; IT, intertrochanter; TR, trochanter; GLCM, gray level cooccurrence matrix; GLSZM, gray level size zone matrix; GLRLM, gray level run length matrix; NGTDM, neighbouring gray tone difference matrix; GLDM, gray level dependence matrix; IDMN, inverse difference moment normalized; SAH GLE, small area high gray level emphasis; GLV, gray level variance; SZN, size zone uniformity; DNN, dependence non-uniformity normalized; GLN, gray level non-uniformity; GLNN, gray level non-uniformity normalized; IMC1, informational measure of correlation 1.

A one-point increase in the bone radiomics score was associated with a 4% elevated risk for hip fracture in multivariable Cox model independent of age, BMI, history of previous fracture, and femoral neck T-score (adjusted HR 1.04, 95% CI 1.02-1.06, p=0.001; Table 5).

연세대학교
YONSEI UNIVERSITY

Table 5. Predictors of incident hip fracture in clinical validation set

	Univariable model		Multivariable model 1		Multivariable model 2	
Predictors of incident hip fracture	Unadjusted HR (95% CI)	P value	Adjusted HR (95% CI)	P value	Adjusted HR (95% CI)	P value
Age (per one year increase)	1.17 (1.09-1.25)	<0.001	1.07 (0.99-1.15)	0.062	1.06 (0.98-1.14)	0.105
Body mass index (per one kg/m^2 increase)	0.91 (0.81-1.02)	0.113	0.97 (0.87-1.09)	0.679	0.98 (0.88-1.11)	0.844
Previous fracture (yes versus no)	5.05 (2.35-10.83)	<0.001	3.06 (1.40-6.70)	0.005	2.86 (1.31-6.23)	0.008
Femoral neck T-score (per one T-score decrease)	3.96 (2.58-6.08)	<0.001	2.73 (1.69-4.42)	<0.001	2.55 (1.54-4.21)	<0.001
Bone radiomics score (per one point increase)	1.06 (1.04-1.09)	<0.001			1.04 (1.02-1.06)	0.001
C-index	Bone radiomics score alone: 0.72		Model 1: 0.80		Model 2: 0.84	
Likelihood ratio χ^2	Bone radiomics score alone: 26.13		Model 1: 54.48		Model 2: 65.22*	

*The bone radiomics score improved the goodness-of-fit of model when added to age, body mass index, previous fracture, and femoral neck T-score by likelihood ratio test (likelihood ratio χ^2 10.74, p=0.001)

Adjustment for osteoporosis as categorical variable instead of femoral neck T-score in multivariable model did not alter the results (adjusted HR 1.04, p<0.001). Further adjustment for hip geometry variables including narrow neck SPW, ECW, CT, BR, or HAL in multivariable Cox model did not attenuate the association between bone radiomics score with hip fracture. The association between bone radiomics score with incident hip fracture remained robust after adjustment for FRAX hip fracture probability score (adjusted HR 1.06 per one bone radiomics score increase, 95% CI 1.03-1.08, p<0.001). Gradient of risk (HR per one SD change) for femoral neck aBMD and bone radiomics score was 3.36 (95% CI 2.30-4.91) and 2.08 (95% CI 1.59-2.72), respectively (Table 6).

Table 6. Gradient of risk (hazard ratio per SD) for femoral neck aBMD and bone radiomics score in clinical validation set

Gradient of risk	Univariate Cox model		Multivariable Cox model*	
	HR/SD (95% CI)	P value	HR/SD (95% CI)	P value
Femoral neck aBMD (per SD decrease)	3.36 (2.30-4.91)	<0.001	2.44 (1.59-3.72)	<0.001
Bone radiomics score (per SD increase)	2.08 (1.59-2.72)	<0.001	1.74 (1.32-2.30)	0.001

*Adjusted for age, BMI, and previous fracture

Gradient of risk for bone radiomics score was adjusted to 1.74 (95% CI 1.32-2.30) in multivariable model including age, BMI, previous fracture (gradient of risk for femoral neck aBMD 2.44 in multivariable model, 95% CI 1.59-3.72). The bone radiomics score improved the model prediction when added to age, body mass index, previous fracture, and femoral neck T-score (likelihood ratio $\chi 2$ 10.74, p=0.001; Table 5). In subgroup analysis, bone radiomics score remained as an independent predictor of incident hip fracture in subgroups by exposure to anti-

연세대학교
YONSEI UNIVERSITY

osteoporosis medication, any falls within a prior year, low handgrip strength, and diabetes mellitus (Table 7).

Table 7. Subgroup analysis in clinical validation set

External test set	Adjusted hazard ratio (95% CI); per one point increase in bone radiomics score*	P value	P for interaction
Anti-osteoporosis medications within five years prior to DXA testing			
Yes (n=487, 24%)	1.05 (1.01-1.10)	0.024	0.617
No (n=1542, 76%)	1.04 (1.01-1.07)	0.011	
Falls within a prior year			
Yes (n=543, 27%)	1.05 (1.01-1.09)	0.019	0.538
No (n=1486, 73%)	1.04 (1.01-1.07)	0.023	
Low handgrip strength (men < 28 kg; women < 18 kg)			
Yes (n=382, 19%)	1.04 (1.00-1.08)	0.047	0.843
No (n=1647, 81%)	1.04 (1.01-1.07)	0.006	
Diabetes mellitus			
Yes (n=464, 23%)	1.05 (1.01-1.11)	0.037	0.214
No (n=1565, 77%)	1.03 (1.01-1.06)	0.012	
CKD (estimated GFR<60 mL/min/1.73m^2)			
Yes (n=664, 33%)	1.02 (0.98-1.06)	0.253	0.215
No (n=1365, 67%)	1.06 (1.03-1.09)	<0.001	

*Adjusted for age, previous fragility fracture, BMI, and femoral neck T-score. Low handgrip strength was defined based on Asian Working Group for Sarcopenia 2019 updated consensus statement (J Am Med Dir Assoc. 2020 Mar;21(3):300-307). Abbreviations: CKD, chronic kidney disease.

Statistical significance of bone radiomics score was attenuated in subgroup with chronic kidney disease (defined as estimated glomerular filtration rate 60 mL/min/1.73m^2 or less) in multivariable Cox model (unadjusted hazard ratio 1.03, 95% CI 1.01-1.08, p=0.013; adjusted hazard ratio 1.02, 95% CI 0.98-1.06, p=0.253). However, interaction between bone radiomics score and presence of chronic kidney disease in multivariable model did not reach statistical significance (p for interaction = 0.215). Association of bone radiomics score with incident hip fracture remained significant (adjusted hazard ratio 1.04, 95% CI 1.01-1.06, p=0.002) after excluding participants with self-reported history of rheumatoid arthritis (n=26, 1.3%), any exposure to systemic glucocorticoid use (n=21, 1.0%), or any self-reported hyperthyroidism (n=31, 1.5%).

연세대학교
YONSEI UNIVERSITY

IV. DISCUSSION

In this study, I found that the bone radiomics score derived from texture features of DXA hip images improved hip fracture risk prediction in community-dwelling older women. The bone radiomics score remained independent predictor of incident hip fracture after adjustment for femoral neck T-score, clinical risk factors, or FRAX hip fracture probability. Bone radiomics score improved predictive performance of the model when added to age, BMI, history of previous fracture, and femoral neck T-score.

Quantitative analysis of bone texture features from various imaging modalities has been actively used to find bone quality markers reflecting bone microarchitecture.[24] The trabecular bone score (TBS) is the current best example for applying the principle of texture analysis to DXA-based vertebral images to quantify the variation in gray-level texture from one pixel to the adjacent pixels.[25,26] Low TBS values indicate lesser gray-level pixel intensity variations in a two-dimensional spine image, which is related to worse bone structure and an elevated risk of major osteoporotic fractures.[25] Given the differences in mechanical, biological characteristics, and the composition of cortical and trabecular bone between the spine and hip bone, textural indices from hip bone images might have distinctive clinical implications, especially for hip fracture risk prediction.[27-29] In a study using 21 human cadaveric hips, texture analysis on excised hip images obtained by a high-resolution x-ray device provided better prediction of the femoral failure load than DXA aBMD alone.[30] A recent study analyzed the spatiotemporal changes of proximal femurs according to age using region free analysis of DXA hip scans from over 13000 Western Europeans, which suggest the potential of texture features based on a pixel-to-pixel relationship in the DXA hip scans as a bone quality biomarker.[13] In line with these findings, I observed that texture features obtained from DXA hip scans showed different patterns between individuals who sustained hip fractures and those who

연세대학교
YONSEI UNIVERSITY

did not. Similar to the findings in TBS, DXA hip scans of individuals who sustained hip fractures had less pixel-to-pixel intensity variation.[25,26] In addition, higher heterogeneity in texture patterns in individuals with hip fracture was observed compared to those without, which may indirectly reflect altered bone microarchitecture of the hip, including cortical bone trabecularization, transformation of trabecular plates to rods, trabecular thinning, and a loss of connectivity leading to bone fragility.[31] As an extension of prior studies, I observed that higher bone radiomics scores estimated using texture features from DXA hip scan images (a model-based summation for the degree of lesser pixel-to-pixel intensity variation and higher heterogeneity in hip bone texture patterns) was associated with elevated risk of hip fracture independent of clinical risk factors and DXA aBMD in older Korean women in a prospective cohort; this had incremental prognostic value for hip fracture when combined with established predictors. Individuals with high risk of fracture had lesser degree of pixel intensity variation (relatively high local homogeneity), whereas they had higher level of texture heterogeneity (abrupt changes in pixel patterns).

DXA is the current standard-of-care imaging modality which guides clinical decisions for detection, initiation of pharmacologic treatment, and follow-up of individuals with osteoporosis and high risk of fracture.[32] Radiomics is an emerging principle for a comprehensive, automated quantification of the radiographic phenotype using data characterization algorithms, which have been intensively studied in oncology fields for tumor characterization and prognostication.[6] In this study, a radiomics-based approach for two-dimensional DXA hip images with machine learning algorithms was able to mine significant texture features related to hip fracture. These findings suggest the potential utility and extensibility of a radiomics-based approach as an add-on procedure to DXA scans in routine clinical practice, including previously stored images, to improve fracture risk prediction. However, there are several challenges in the radiomics modeling process, including susceptible data reproducibility, particularly

연세대학교
YONSEI UNIVERSITY

regarding segmentation, high dimensionality (more features than observations) leading to overfitting, and inherent high correlation among features.[33,34] In this study, segmentation of ROIs guided by contour lines in DXA hip images enabled relatively robust inter-rater reproducibility (ICC>0.90) in most of the texture features from each ROI and in all of 14 features selected in the final model. To reduce dimensionality and to avoid collinearity, I applied various machine learning principles to select as few features as possible with the optimal predictive performance. To test the generalizability of the bone radiomics score, analysis was performed using a community-based prospective cohort dataset with incident fracture data. Although further meticulous studies need to be performed to validate the utility of the bone radiomics score, our study showed the potential of a radiomics approach for DXA images to improve fracture prediction along with DXA aBMD.

In a prior study performed in UK, obese group (BMI 30 kg/m^2 or higher) had higher precision error (%CV) in femoral neck BMD compared to normal (BMI < 25 kg/m^2) or overweight group (BMI 25-29.9 kg/m^2), partly due to increased soft tissue thickness and inhomogeneity.[35] However, difference of precision error in femoral neck BMD or total hip BMD between normal and overweight group did not reach statistical significance. In our derivation set and clinical validation cohort, most of subjects had BMI within normal to overweight range (96% and 95%; mean BMI 23.4 kg/m^2 and 24.5 kg/m^2), whereas 4% and 5% had BMI at obese range (BMI 30 kg/m^2 or higher). Among all radiomics features, GLRLM gray-level non-uniformity (total hip, r=0.37; intertrochanter, r=0.28; trochanter, r=0.24, p<0.001 for all) and NGTDM coarseness (total hip r=-0.30; intertrochanter r=-0.20; p<0.001 for all) showed weak to moderate correlation with weight, all of which were not included in the final model. Other radiomics features showed weak to negligible correlation (-0.2 to 0.2) with weight or BMI. Although I did not observe significant interference by weight on bone radiomics score in this study, it remains unclear whether radiomics features

연세대학교
YONSEI UNIVERSITY

would be affected by weight in obese individuals with BMI 30 kg/m^2 or higher, which merits further investigation.

This study has several limitations. Although I tried to assess hip fracture-free status in KURE cohort by interviewer-assisted questionnaires during follow-up with further ascertainment of the absence of any claims for ICD10 codes (S72.0, S72.1; hip fracture codes) based on data linkage with Korean HIRA database covering 98% of entire Korean population, I was not able to directly confirm fracture-free status at individual level by reviewing medical records and imaging data. This may have affected the results by classifying some individuals at high risk as low risk group. However, the incidence rate of hip fracture in older women participants of KURE cohort in this study was 334.6/100,000 person-year, which was similar to previously reported incidence rates in Korean older women with similar age (age 75-79, 332.8/100,000 person-year) using nationwide database.[36] Hip fracture cases for the derivation set were retrieved from the Severance Hospital EHR database while the age-matched subset of KURE cohort was used to provide non-fracture controls for the derivation set. Although excluding the subset used for derivation set from clinical validation set was inevitable to maintain integrity of clinical validation set and the number of subset used for derivation set was small relative to entire cohort size, it might bias the underlying risk assessment by systematically excluding those who did not have fracture. The radiomics model was developed to capture the difference in textural characteristics of DXA images of individuals who sustained a hip fracture or not; the scope of the outcome was limited to hip fracture in this study. AUPRC in test set was lower than AUROC, which may relate to class imbalance.[37] Because our data did not have TBS measurements, I could not analyze the correlation between TBS and the bone radiomics score. The bone radiomics score may have incremental value for DXA aBMD and TBS and this needs to be examined in future studies. Statistical non-significance of age in multivariable models may reflect the relatively small size of the cohort with limited power. Although this study is a proof-of-concept study, I acknowledge that number of outcomes are limited in both derivation set and clinical validation set. Further investigation in larger cohort is needed to examine the incremental prognostic value of bone radiomics score over BMD or clinical risk factors. Our findings cannot be generalized to ethnicities other than Koreans or men at this stage. As our data are based on a single DXA manufacturer (Hologic), further studies are needed to apply the bone radiomics score model for hip images obtained from other DXA manufacturers. Information about anti-osteoporosis medication use after DXA testing during follow-up were not available in this study. Although a prior study showed the association between texture features from proximal femur and femoral strength using human cadaveric femurs, the biologic relevance of the bone radiomics score needs to be further explored.[30] Although all analysis procedures were semi-automated, automation of segmentation and mask generation process would facilitate effective application of radiomics-based approach to DXA images. Textural heterogeneity in the cortical and trabecular bone in DXA hip images could not be discerned due to limited resolution of DXA images.

V. CONCLUSION

In conclusion, the bone radiomics score based on texture features from DXA hip images was associated with an elevated risk of hip fracture, independent of clinical risk factors and the DXA T-score. The bone radiomics score may have the potential to improve fracture risk prediction as an addition to DXA testing in current standard-of-care practice; this needs to be validated in further studies.

연세대학교
YONSEI UNIVERSITY

References

1. Kanis JA, Odén A, McCloskey EV, Johansson H, Wahl DA, Cooper C, et al. A systematic review of hip fracture incidence and probability of fracture worldwide. Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 2012;23:2239-56.
2. Curtis EM, Moon RJ, Harvey NC, Cooper C. The impact of fragility fracture and approaches to osteoporosis risk assessment worldwide. Bone 2017;104:29-38.
3. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015;386:743-800.
4. Kanis JA, Kanis JA. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Synopsis of a WHO report. Osteoporosis International 1994;4:368-81.
5. Schuit SCE, van der Klift M, Weel AEAM, de Laet CEDH, Burger H, Seeman E, et al. Fracture incidence and association with bone mineral density in elderly men and women: the Rotterdam Study. Bone 2004;34:195-202.
6. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research 2017;77:e104.
7. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.
8. Rastegar S, Vaziri M, Qasempour Y, Akhash MR, Abdalvand N, Shiri I, et al. Radiomics for classification of bone mineral loss: A machine learning study. Diagn Interv Imaging 2020;101:599-610.

연세대학교
YONSEI UNIVERSITY

9. Amling M, Herden S, Posl M, Hahn M, Ritzel H, Delling G. Heterogeneity of the skeleton: comparison of the trabecular microarchitecture of the spine, the iliac crest, the femur, and the calcaneus. J Bone Miner Res 1996;11:36-45.

10. Carballido-Gamio J, Harnish R, Saeed I, Streeper T, Sigurdsson S, Amin S, et al. Proximal femoral density distribution and structure in relation to age and hip fracture risk in women. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 2013;28:537-46.

11. Yu A, Carballido-Gamio J, Wang L, Lang TF, Su Y, Wu X, et al. Spatial Differences in the Distribution of Bone Between Femoral Neck and Trochanteric Fractures. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 2017;32:1672-80.

12. Yamamoto N, Sukegawa S, Kitamura A, Goto R, Noda T, Nakano K, et al. Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates. 2020;10:1534.

13. Farzi M, Pozo JM, McCloskey E, Eastell R, Harvey N, Wilkinson JM, et al. A Spatio-Temporal Ageing Atlas of the Proximal Femur. IEEE transactions on medical imaging 2020;39:1359-68.

14. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30:1323-41.

15. Aye Aye A, Zin Mar W. Preprocessing with Contrast Enhancement Methods in Bone Age Assessment. In: Lee R, editor. Computer and Information Science. Cham: Springer International Publishing; 2020. p.31-45.

16. Ikhsan IAM, Hussain A, Zulkifley MA, Tahir NM, Mustapha A. An analysis of x-ray image enhancement methods for vertebral bone

연세대학교
YONSEI UNIVERSITY

segmentation. 2014 IEEE 10th International Colloquium on Signal Processing and its Applications; 2014. p.208-11.

17. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-e7.

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. 2012. p.arXiv:1201.0490.

19. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review 1950;78:1-3.

20. Murphy AH. A Note on the Ranked Probability Score. Journal of Applied Meteorology 1971;10:155-6.

21. Cochran WG. Some Methods for Strengthening the Common χ<sup>2</sup> Tests. Biometrics 1954;10:417-51.

22. Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. Biometrics 1983;39:499-503.

23. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806-8.

24. GENANT HK, JIANG Y. Advanced Imaging Assessment of Bone Quality. 2006;1068:410-28.

25. Silva BC, Leslie WD, Resch H, Lamy O, Lesnyak O, Binkley N, et al. Trabecular bone score: a noninvasive analytical method based upon the DXA image. J Bone Miner Res 2014;29:518-30.

26. Hans D, Goertzen AL, Krieg M-A, Leslie WD. Bone microarchitecture assessed by TBS predicts osteoporotic fractures independent of bone density: The manitoba study. 2011;26:2762-9.

27. Cauley JA, Blackwell T, Zmuda JM, Fullman RL, Ensrud KE, Stone KL, et al. Correlates of trabecular and cortical volumetric bone mineral

density at the femoral neck and lumbar spine: the osteoporotic fractures in men study (MrOS). Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 2010;25:1958-71.

28. Johansson H, Kanis JA, Odén A, Leslie WD, Fujiwara S, Glüer CC, et al. Impact of femoral neck and lumbar spine BMD discordances on FRAX probabilities in women: a meta-analysis of international cohorts. Calcified tissue international 2014;95:428-35.

29. Amling M, Herden S, Pösl M, Hahn M, Ritzel H, Delling G. Heterogeneity of the skeleton: comparison of the trabecular microarchitecture of the spine, the iliac crest, the femur, and the calcaneus. J Bone Miner Res 1996;11:36-45.

30. Le Corroller T, Halgrin J, Pithioux M, Guenoun D, Chabrand P, Champsaur P. Combination of texture analysis and bone mineral density improves the prediction of fracture load in human femurs. Osteoporos Int 2012;23:163-9.

31. Chavassieux P, Seeman E, Delmas PD. Insights into Material and Structural Basis of Bone Fragility from Diseases Associated with Fractures: How Determinants of the Biomechanical Properties of Bone Are Compromised by Disease. Endocrine Reviews 2007;28:151-64.

32. Schousboe JT, Shepherd JA, Bilezikian JP, Baim S. Executive summary of the 2013 International Society for Clinical Densitometry Position Development Conference on bone densitometry. J Clin Densitom 2013;16:455-66.

33. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. Korean J Radiol 2019;20:1124-37.

34. Younes Qasempour MRA, Isaac Shiri, Ghasem Hajianfar, Neda Abdalvand, H Abdollahi. Test–Retest Reproducibility Analysis of Bone

연세대학교
YONSEI UNIVERSITY

Mineral Densitometry Radiomics Features, 27 July 2020, PREPRINT (Version 1) available at Research Square. doi:10.21203/rs.3.rs-44885/v1.

35. Knapp KM, Welsman JR, Hopkins SJ, Fogelman I, Blake GM. Obesity Increases Precision Errors in Dual-Energy X-Ray Absorptiometry Measurements. Journal of Clinical Densitometry 2012;15:315-9.

36. Cheung C-L, Ang SB, Chadha M, Chow ES-L, Chung Y-S, Hew FL, et al. An updated hip fracture projection in Asia: The Asian Federation of Osteoporosis Societies study. Osteoporosis and sarcopenia 2018;4:16-21.

37. Ozenne B, Subtil F, Maucort-Boulch D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015;68:855-9.

국문요약

골밀도 영상 라디오믹스 텍스처 분석 및 머신러닝 기반 대퇴골절 예측모델 수립

<지도교수 이유미>

연세대학교 대학원 의학과

홍남기

내 용

이중에너지 X-선 흡수계측법 (DXA) 골밀도검사는 골절고위험군을 판별하기 위한 골다공증 진단의 표준검사이다. 대퇴부 골밀도 영상을 이용한 라디오믹스 분석은 추가 비용 없이 골절예측력을 개선할 가능성이 있다. 본 연구는 지역사회 거주 노인 여성에서 대퇴골밀도 영상 기반 라디오믹스 대퇴골절 예측모델 (골라디오믹스점수) 을 구축하고, 골라디오믹스점수가 대퇴골절 예측력 개선 여부를 검정하였다. 모델구축코호트은 세브란스병원 후향코호트를 기반으로 구성되었다 (140명, 골밀도 기반조사 후 추적관찰기간 중 대퇴골절 경험, 평균나이 73세, 골절까지 기간 중위수 2.1년; 290명, 나이 매칭한 추적관찰 기간 중 대퇴골절을 경험하지 않은 대조군, 평균나이 73세, 추적관찰기간 중위수 5.5년). 모델구축코호트는 75%의 학습데이터셋과 25%의 테스트데이터셋으로 나뉘었다. 대퇴골밀도 영상에서 얻어진 300개의 라디오믹스 특성변수 중 14개의 최적 특성변수를 기계학습기법을 통해 추출하였고, 이를 이용해 예측모델을 구축하였을 때 랜덤포레스트 기반 모델이 테스트셋에서 최적 성능을 보여주었고 이를 토대로 골라디오믹스점수를 계산하였다 (0-100점). 임상검증데이터셋인 지역사회코호트에서 (2029명 여성, 평균나이 71세), 앞서 구축한 학습데이터셋에서 학습한 모델을 기반으로 골라디오믹스점수를 계산하였다. 총 34명 (1.7%)의 여성에서 중위수 5.4년 추적관찰 기간 동안 대퇴골절이 발생하였다 (평균 골라디오믹스점수 골절군 40±16 대 비골절군 28±12 점, $p<0.001$). 골라디오믹스가 1점 상승할 때 나이, 체질량지수, 기존골절력, 대퇴골밀도 T점수와 독립적으로

연세대학교
YONSEI UNIVERSITY

대퇴골절발생위험도가 4% 증가하였으며 (보정 위험비 [aHR] 1.04, p=0.001), 골라디오믹스점수는 기존 예측인자로 이루어진 기본모델에 추가되었을 때 모델의 적합도를 유의하게 개선시켰다 (likelihood ratio χ^2 10.74, p=0.001). 골라디오믹스점수의 대퇴골절 예측력은 기존 골절예측모델은 FRAX 점수를 보정하였을 때도 독립적으로 유의하였다 (aHR 1.06, p<0.001). 대퇴골밀도 영상에서 추출한 골라디오믹스점수는 추가 비용없이 대퇴골절 예측력을 개선시킬 가능성이 있다.

핵심되는 말 : 이중 에너지 X-선 흡수계측법, 골절위험도예측, 골다공증

연세대학교
YONSEI UNIVERSITY

PUBLICATION LIST

Hong, N., Park, H., Kim, C.O., Kim, H.C., Choi, J.-Y., Kim, H. and Rhee, Y. Bone Radiomics Score Derived From DXA Hip Images Enhances Hip Fracture Prediction in Older Women. J Bone Miner Res. https://doi.org/10.1002/jbmr.4342