creative commons
COMMONS DEED

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

Language guided image generation to enhance fracture risk prediction using lateral spine plain radiograph

Sang Wouk Cho

Department of Integrative Medicine

The Graduate School, Yonsei University

연세대학교
YONSEI UNIVERSITY

Language guided image generation to enhance fracture risk prediction using lateral spine plain radiograph

A Dissertation Submitted to the

Department of Integrative medicine

and the Graduate School of Yonsei University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Health Engineering

Sang Wouk Cho

December 2023

연세대학교
YONSEI UNIVERSITY

This certifies that the dissertation of
Sang Wouk Cho is approved.

Thesis Supervisor: Namki Hong

Hwiyoung Kim: Thesis Committee Member #1

Kyoung Min Kim: Thesis Committee Member #2

Seng Chan You: Thesis Committee Member #3

Kyu-Hwan Jung: Thesis Committee Member #4

The Graduate School
Yonsei University
December 2023

연세대학교
YONSEI UNIVERSITY

Acknowledgements

This page is specifically designed to express my profound gratitude and respect for those who supported the completion of my thesis. I am deeply indebted to my supervisors and mentors, Professor Namki Hong and Hwiyoung Kim, for their unwavering encouragement, help, guidance, and support throughout this study.

I would like to express my sincere gratitude to thesis committee members, Professor Kyoung Min Kim, Seng Chan Yu, and Kyu-Hwan Jung who provided fundamental insights to improve this study. I especially thank Sookyeong Han, Minkyu Kim, Byounghan Jang and Teayun Park for their efforts in gathering dataset and research discussion. I am also indebted to all members of Yonsei Bone and Mineral Research team and Translational Artificial Intelligence Laboratory.

My family deserves endless gratitude: my wife Taeeun Kim for being the love of my life; my parents for their love and wisdom that guided me; my sisters, whose unwavering support and encouragement have been a constant source of strength and joy throughout this journey; and my grandparents, whose stories, lessons, and unconditional love have deeply enriched my life and work. Their belief in my abilities has fueled my determination, and I am forever grateful for their unwavering faith in me.

In closing, I reflect upon the journey that this thesis has been and the invaluable lessons learned along the way. I am deeply grateful to everyone who supported me, both seen and unseen, during this endeavor.

연세대학교
YONSEI UNIVERSITY

TABLE OF CONTENTS

연세대학교
YONSEI UNIVERSITY

LIST OF TABLES

연세대학교
YONSEI UNIVERSITY

LIST OF FIGURES

연세대학교
YONSEI UNIVERSITY

ABSTRACT

Language guided image generation

to enhance fracture risk prediction using lateral spine plain radiograph

Sang Wouk Cho

Department of Integrative Medicine

The Graduate School of Yonsei University

Directed by Professor Namki Hong and Hwiyoung Kim

Spine radiography along with deep neural networks is capable of detecting prevalent vertebral fractures and osteoporosis. However, whether the generative model predicts fracture risk remains uninvestigated.

Clinical variables and lateral spine X-ray images from patients aged 50 or older who presented to Severance Hospital, Korea between January 2007 and December 2018 were collected. The incident fracture was defined using follow-up X-ray radiographs. Our model consists of two language-guided latent diffusion models (LDM) to extract feature maps of morphological structure and generate new images with clinical prompts on training set (80% hold-out set) and test set (20% hold-out set). Verte-X prevalent vertebral fracture scores (pVF scores) were calculated on the baseline images (BpVF) and 10-year generative images (GpVF).

연세대학교
YONSEI UNIVERSITY

Fracture risk assessments were then conducted, categorizing them into three groups based on these scores: 1) low risk for both BpVF & GpVF (LL), 2) low risk for BpVF & high risk for GpVF (LH), 3) and high risk for BpVF regardless of GpVF status.low risk for both BpVF and GpVF (LL), low risk for BpVF but high for GpVF (LH), and high risk for BpVF regardless of GpVF status.

A total of 29,307 lateral spine plain X-rays for 9,276 patients with (mean age 65.7 years, women 66%; VF prevalence 18.6%) were analyzed in the derived cohort. Over a mean follow-up period of 34.8 months, 9.9% of patients experienced vertebral fractures (921 out of 9,276 in the whole dataset) after baseline. Generative images revealed possible changes in the spine at different time points. The mean (SD) error in pVF scores between real-follow up and generative X-ray images was 0.06 ± 0.20 with a correlation coefficient r of 0.655 (0.547,0.741). When stratified into the risk group, LH group and HH risk group were associated with 109% and 391% increased risk of fracture respectively (hazard ratio [HR], 2.092 and 4.911; P=<0.001 for all), showing an improved model fit by adding age, sex, and BMI to covariates (likelihood ratio 105.7, p <0.001). The association between risk groups with incident fracture remained robust ([HR] 1.461; P=<0.001) after adjustment for FRAX major osteoporotic probability.

In summary, generative image-based risk stratification showed its potential to improve clinical workflow and fracture risk assessments.

Key words: Osteoporosis, Fracture risk, Deep learning, Generative AI, Survival analysis

연세대학교
YONSEI UNIVERSITY

Language guided image generation
to enhance fracture risk prediction using lateral spine plain radiograph

Sang Wouk Cho

the Department of Integrative Medicine
the Graduate School of Yonsei University

Directed by Professor Namki Hong and Hwiyoung Kim

I. INTRODUCTION

1. Background

Osteoporotic fractures (OF) are fractures that have a high mortality and morbidity rate, aggravating individual health and the economy around the world. [1-3] 8.9 million new fractures are caused by osteoporosis around the world indicating that a new OF occurs every 3 seconds. [4] Therefore, the prevention of osteoporotic vulnerable fractures is a big challenge for the growing elderly population. [5] Despite the high efficiency of medication to reduce fracture risk, patients with osteoporosis and morphological fractures who are in high risk for future fractures are still underdiagnosed. [6]

Measurement of the areal bone mineral density (BMD) using dual-energy x-ray absorptiometry (DXA) is the gold standard for diagnosing osteoporosis and for

연세대학교
YONSEI UNIVERSITY

evaluating bone strength in clinical practice. However, many fractures could occur at osteopenia [9] which means BMD is not a reliable factor for fracture risk prediction. [10, 11] The fracture risk assessment tool (FRAX) is one of the fracture risk evaluation tools using clinical variables with BMD at the femoral neck. [12-16] As FRAX is the fracture risk prediction tool, the FRAX major osteoporotic and hip fracture scores are integrated into risk stratification guidelines of clinical practice in South Korea and the United Kingdom. [17, 18] When calculating FRAX scores, some of the factors known to be associated with fracture risk are excluded, such as increment of fracture risk after initial fracture. [19, 20]

Medical images are important resources that provide evidence for diagnosis and clinical decisions. [21] Diverse modality of medical images may include potential biomarkers to predict patients' prognosis [22], also providing diagnostic information to understand their medical issues. Recent successful deep learning technology has been applied to medical imaging [23-25] such as Generative Adversarial Networks (GANs). [26] As of today, diffusion models have gained attention in the field of generative model due to their ability to generate impressive quality images. [27] During the diffusion process, they iteratively add Gaussian Noise in the noising process and learn to denoise to generate sample image. Another advantage of the diffusion model is conditional image generation with guidance such as a sentence [28] and an image mask [29]. Conditioned models allow diverse and customized results with image editing and translation.

Spine X-ray image is a widely available medical imaging modality in clinical practice which provides an amount of information, including bone density, morphological spine structure, and soft tissues. Especially, detection of prevalent

연세대학교
YONSEI UNIVERSITY

vertebral fractures using lateral spine X-ray images is recommended by the International Society for Clinical Densitometry (ISCD) guideline. [30] The broad applicability and rich information content of spine X-ray images make this modality a valuable resource for developing artificial intelligence models capable of predicting individual fracture risks. Additionally, stratifying incident fracture risk using spine X-ray images has the potential to facilitate preventive interventions in various risk groups, thereby mitigating fragility fractures and their associated negative consequences.

In the field of fracture risk prediction, prior studies focused on building multi-variable fracture risk models to predict individual risk scores. [12-16, 31-34] In early proposed studies [12-16], BMD was considered the primary determinant, and recently, studies[31-34] using images have been presented. X-ray images have received less attention for fracture risk prediction compared to DXA, computed tomography (CT), or magnetic resonance imaging (MRI) [35]. Hsieh and colleagues showed that convolutional neural networks could assess fracture risk using pelvis/lumbar spine radiographs involving 5,164 and 18,175 patients, respectively. [32] Additionally, Kong and colleagues demonstrated that a CNN-based survival prediction algorithm, utilizing baseline images and clinical variables, outperformed the FRAX and CoxPH models in the assessment of lumbar spine radiographs. [31] Two studies have shown that X-ray images can predict the risk of incident fractures, indicating that X-ray images contain diagnostic information for understanding fracture development. However, these studies had limitations related to using only the lumbar spine view position, and the potential for generating follow-up X-ray images has not yet been explored. Furthermore, it remains unclear whether risk stratification using personalized

연세대학교
YONSEI UNIVERSITY

quantified scores from generative models would lead to improvements in clinical interventions for incident fracture risk.

In our previous study [36], we devised deep learning scores for osteoporosis detection and the identification of prevalent vertebral fractures using lateral spine radiography, denoted as the 'Verte-X osteo score' and 'Verte-X pVF score.' We also explored the potential of these scores for enhancing the referral of high-risk individuals for bone-density testing.

In this study, we introduce a language-guided diffusion model designed to generate follow-up X-ray images of patients over aging and calculate personalized Verte-X pVF scores based on both baseline images and generative follow-up images. Our aim is to investigate whether these individualized scores, derived from spine X-ray images, can stratify patients into incident fracture risk groups, ultimately improving the clinical approach to fracture risk assessment and achieving patient-centered care.

2. Technical related work

Image synthesis across medical images is a very active field to facilitate clinical procedures and generate images of rare diseases. [37] Particularly, Generative Adversarial Networks (GAN) [38] had been the most popular method to solve those problems. Since the publication of Denoising Diffusion Probabilistic Models [39] in 2020, GAN has been replaced gradually by diffusion models. The performance of the diffusion-based model was much better and more stable than that of GAN in abnormal detection [40-42], multi-modality translation [43], and meta-data generation tasks. [44] Additionally, images could be generated by

연세대학교
YONSEI UNIVERSITY

conditions or text using guided diffusion models, eliminating the collapse issues of GAN.[45]

A recent publication [46] has demonstrated that the latent space within the U-net, computed at each timestep of the Diffusion model, encompasses morphological information derived from the original image. In addition, another study[47] demonstrates that combining the feature from the Resnet block, the key and query utilized in Self-attention during the up-sampling process of the U-net could give the original image a Text guidance effect. This means that if a certain condition is provided in the model given as input, it could be translated to the image closest to the condition. However, both studies were tested with object-oriented target examples such as humans, animals, or buildings. Additional research is required to validate their applicability in complex medical images, and to assess the continued utility of these methodologies.

연세대학교
YONSEI UNIVERSITY

II. MATERIALS AND METHODS

1. Study subjects

The language-guided future image generation model dataset – the VERTEbral fracture and osteoporosis detection in spine X-ray [VERTE-X] cohort was previously described in our work. [36] Data use and study concept were approved by the Institutional Review Board of Severance Hospital, Seoul, Korea, with the waiver of written informed consent for medical-record review (IRB no. 4-2021-0937). The VERTE-X cohort comprises individuals who received lateral spine X-ray examination at Severance Hospital, Seoul, Korea, between January 2007 and December 2018.

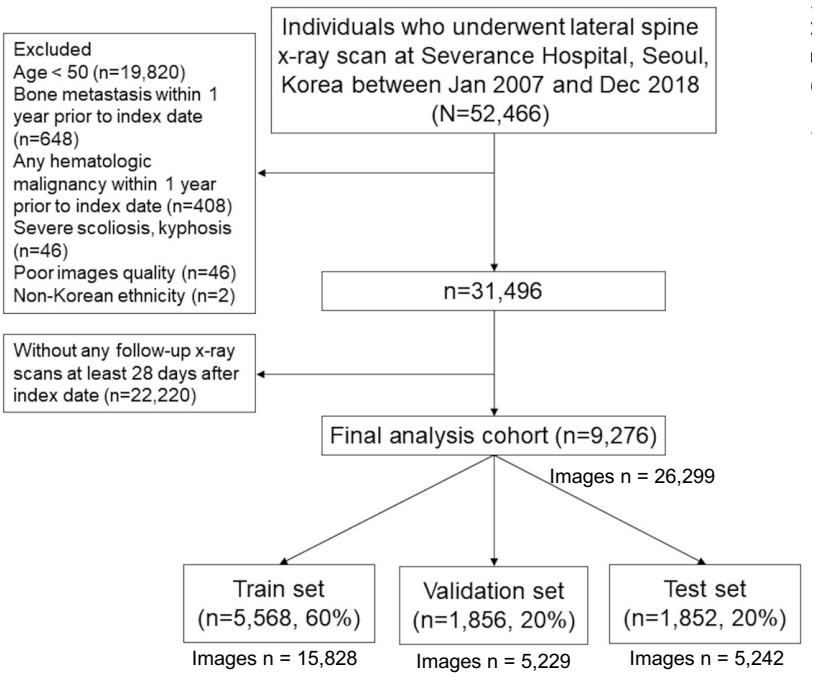


Figure 1 Study flow

연세대학교
YONSEI UNIVERSITY

To improve the data quality and extend the utility of the cohort for fracture studies, we refined the initial dataset, which included over 140,000 spine X-ray images from 52,466 individuals, based on the exclusion criteria outlined in Supplemental Figure 1. To specifically target older individuals with high prevalence and clinical relevance to vertebral fractures or osteoporosis, those under the age of 50 were excluded (n=19,820). Exclusion criteria encompassed individuals of non-Korean ethnicity (n=2), those with a history of bone metastasis or hematologic malignancy within one year preceding the index date (n=648), as those with severe scoliosis, kyphosis (n=46), or missing DICOM files (n=46). To assess the clinical validation of future incident fractures, we retained 9,276 individuals with 26,299 lateral spine X-ray images, all of whom underwent X-ray examination at least 28 days after the index date in the final derivation cohort. They were randomly divided into three groups: model development (n = 5,568, 60%), model validation (n = 1,856, 20%), and testing (a hold-out set, n = 1,852, 20%) while maintaining the proportions of age groups, sex and outcome prevalence.

2. Input features

1) X-ray images: Lateral spine X-ray images in the derivation set were obtained using machines from 12 different manufacturers (Supplemental Table 1).
2) Clinical features: Clinical features of study participants were retrieved using the Severance Hospital Clinical Data Warehouse (SCRAP 2.0) system. Demographic information, including age, sex, weight, and height, was collected. In addition, we collected data on previous clinical fractures, the presence of rheumatoid arthritis, and certain causes of secondary osteoporosis (such as malabsorption, end-stage liver disease, and type-1 diabetes mellitus) using relevant ICD-10 diagnosis codes,

연세대학교
YONSEI UNIVERSITY

which were ascertained through medical record reviews. We also defined glucocorticoid use as any patient who had been prescribed a daily dose equivalent to prednisolone 5 mg or higher for at least three times of the previous six months. FRAX major osteoporotic score was calculated using the University of Sheffield's online FRAX tool (https://frax.shef.ac.uk/FRAX/tool.aspx). During the FRAX score calculation, characteristics of age, sex, BMI, previous fracture, parent hip fracture, current smoking, use of glucocorticoids, rheumatoid arthritis, secondary osteoporosis, alcohol intake 3 or more units/day, and femoral neck BMD were used.

3. Definition of outcomes

After retrieving digital images of lateral spine radiographs from the Picture Archiving and Communication System (PACS), the presence of vertebral fractures was determined using an algorithm-based qualitative method based on the lateral spine X-ray images. [48, 49]

연세대학교
YONSEI UNIVERSITY

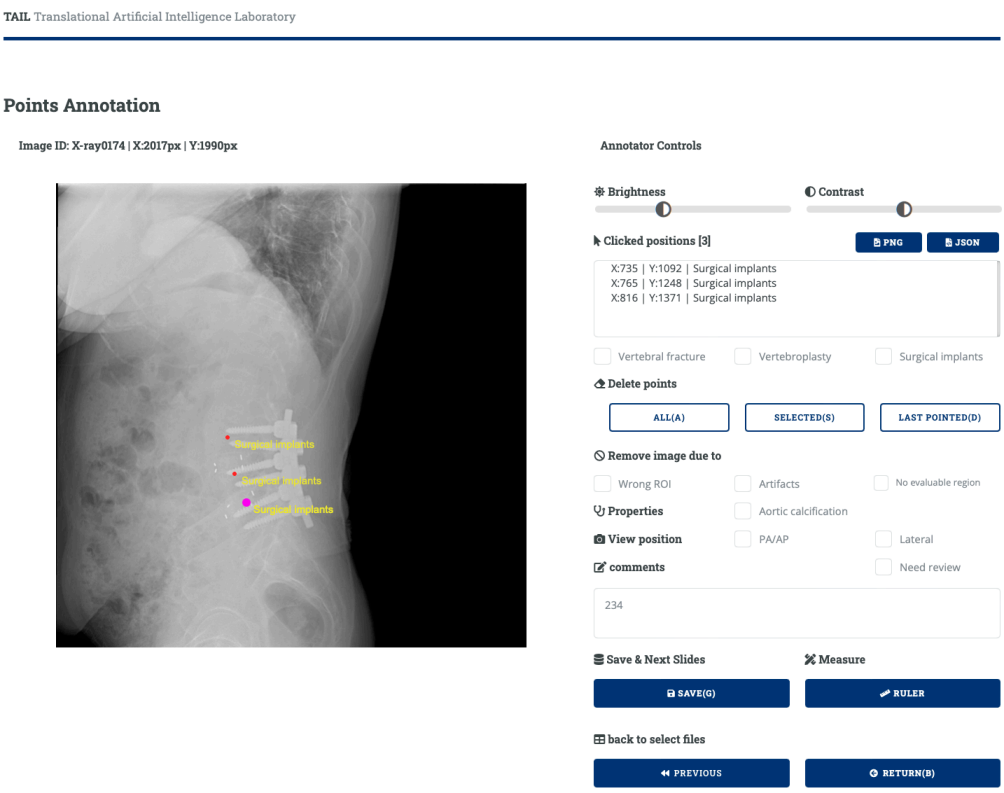


Figure 2 Annotation tools for spine vertebral fracture

A clinical annotation system was developed for identifying the presence of vertebral fractures in lateral spine radiographic images and curated by four independent reviewers with five years of clinical practice experience. Discrepancies in curated labels were checked by two expert reviewers with over ten years of clinical practice experience. In a subset of participants who had available DXA results within one year of the index date, osteoporosis was defined as a DXA-derived T-score $\leq$2.5 at the lumbar spine, femoral neck, or total hip, referencing NHANES III Caucasian young female mean and standard deviation.

연세대학교
YONSEI UNIVERSITY

[50] Two DXA machines, Discovery W and Discovery A (Hologic, MA, USA), were used during the study period on the derivation cohort.

4. Image processing

For VERTE-X pVF score calculation, DICOM files of the lateral spine X-ray images were downloaded from PACS and resized to 1024*512 pixels while preserving the original width and height aspect ratio. Contrast-Limited Adaptive Histogram Equalization (CLAHE) and normalization were applied to the lateral spine X-ray images. Detailed information on image processing is provided below the figure.

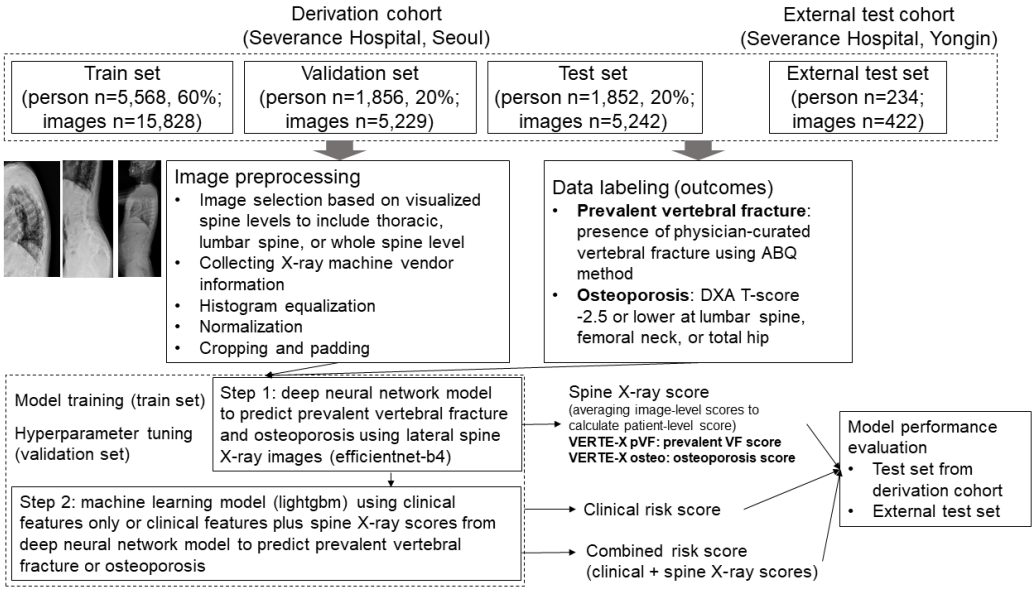


Figure 3 Development and validation process of spine x-ray score to predict prevalent vertebral fracture and osteoporosis using machine learning algorithms

In preparation for future image generation, X-ray images were resized to 512x512 pixels and paired with relevant clinical characteristics, including age, sex, BMI,

연세대학교
YONSEI UNIVERSITY

the prevalence of vertebral fracture in the current image, glucocorticoid use, the history of previous clinical fractures, the presence of rheumatoid arthritis, and the occurrence of secondary osteoporosis. These attributes were used to generate prompts for describing the image status of each individual image.

5. Prevalent fracture score (Verte-X pVF score)

In a previous study[36], two separate deep convolutional neural network (DCNN) models (EfficientNet-B4)[51] were trained based on lateral spine radiography to detect prevalent vertebral fractures (VERTE-X pVF score) and the presence of osteoporosis (VERTE-X osteo score). After model optimization, scores for each outcome (ranging from 0 to 1) were obtained from the output layer for each spine X-ray image input.

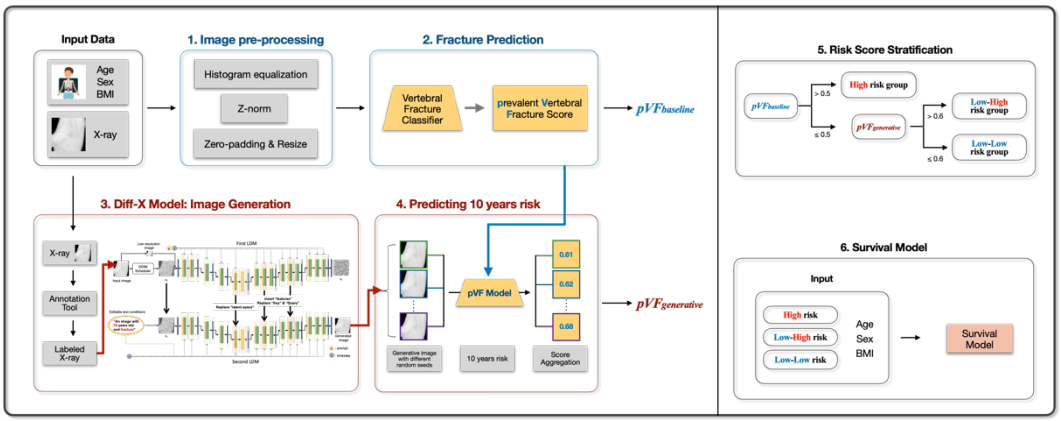


Figure 4 Work flow

6. Diff-X: language-translated X-ray generation model

In this study, we developed a diffusion-based X-ray image generation model using a language prompt-guided module named Diff-X. The Diff-X model is a

연세대학교
YONSEI UNIVERSITY

modification of the Play-and-Plug model [47], which takes text-to-image synthesis to the realm of image-to-image translation. This modification aimed to enhance the preservation of the original image structure and extract more accurate spine structure features.

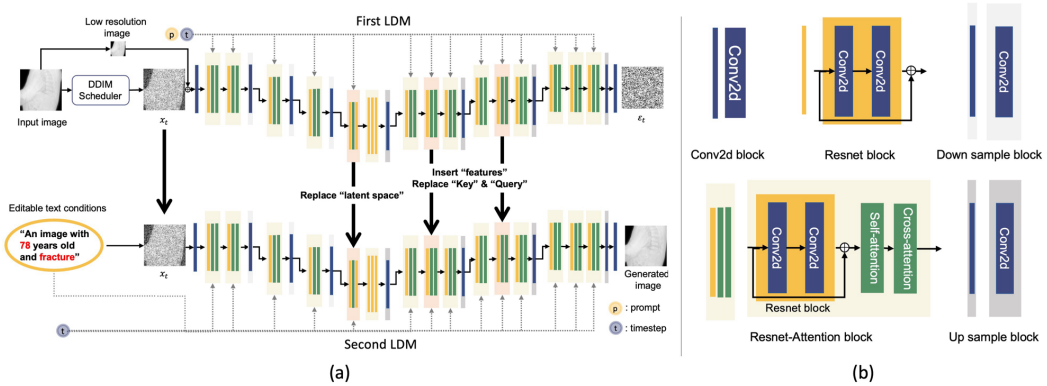


Figure 5 Diff-X model structure and blocks. The model consists of two distinct LDMs: The first LDM extracts structure information and the second LDM generates age-translated images

We trained two distinct Latent Diffusion Models (LDMs, here we used a Stable diffusion structure) [28] as shown in the model structure; one for extracting the morphological structure of the input image and another for giving guidance on prompts. The first LDM extracts features, keys, and queries from the input image and this latent space is added to the second LDM, which then generates an age-predicted image.

To develop robust reconstructions on the first LDM, we followed the previous study[52]. The low-resolution version of the input image is concatenated channel-wise with the input image after the DDIM (Denoising diffusion implicit models) Scheduler for every t timestep. This process provides the original information to

연세대학교
YONSEI UNIVERSITY

generate an image closer to the input image. The U-net blocks consist of three Resnet blocks and nine Resnet-attention blocks. Narek Tumanyan's study[47] attempted to add the features, keys, and queries of the fourth layer of the Resnet-attention block. However, we empirically found that the fifth block contains more information in our case. Kwon's study[46] showed that manipulating the latent space, called h-space, could edit diverse styles of images. We replaced the latent space, key, and query of the second LDM with those of the first LDM at the same location for every t timestep. Finally, the reconstruction feature is adopted to the text-guided age feature map and generates age-dependent images.

Additionally, we showed the difference in the age-translated images to inspect the actual changes in age. The differences tell which characteristic appears as age changes. We also demonstrate how comprehensive the feature extraction contains spatial information. Tumanyan, N's study [47] showed PCA analysis could visualize the most important components in RGB channels. We present all the features from each layer. Feature maps and examples of the future generative images are depicted below.

연세대학교
YONSEI UNIVERSITY

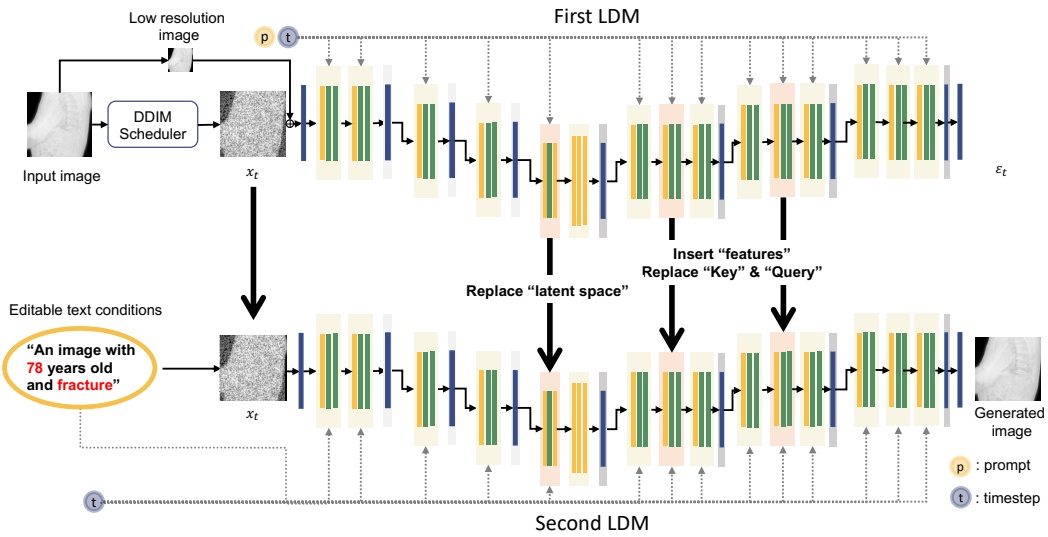


Figure 6 Model architecture of the Diff-X model

7. Score generation and risk stratification

1) Fracture score calculation

Using the Diff-X model, future X-ray images were generated from the baseline image and the clinical prompt at 1 year, 5 years, 10 years, and the last follow-up duration. At each time point, these images were generated 10 times using different random seeds. While a constant seed allowed for reproducibility and the preservation of certain parts of the image across the prompt given [53], this approach could achieve answer coverage through repeated measurements with different random seeds. [54-56] In this study, using different random seeds enabled the model to retain the core information from the clinical prompt and baseline images, resulting in a range of possible future images. VERTE-X pVF scores were calculated for both baseline images and each generative image. For patient-level aggregation, future images generated from the same baseline image were averaged, and the maximum value within the same participant was then

연세대학교
YONSEI UNIVERSITY

selected. These scores from the baseline image and generative image were named 'Baseline pVF score (BpVF)' and 'Generative pVF score (GpVF)', respectively. The uncertainty scores were calculated based on the variation of GpVF to reveal the characteristics of individuals and improve the reliability of generative risk scores.

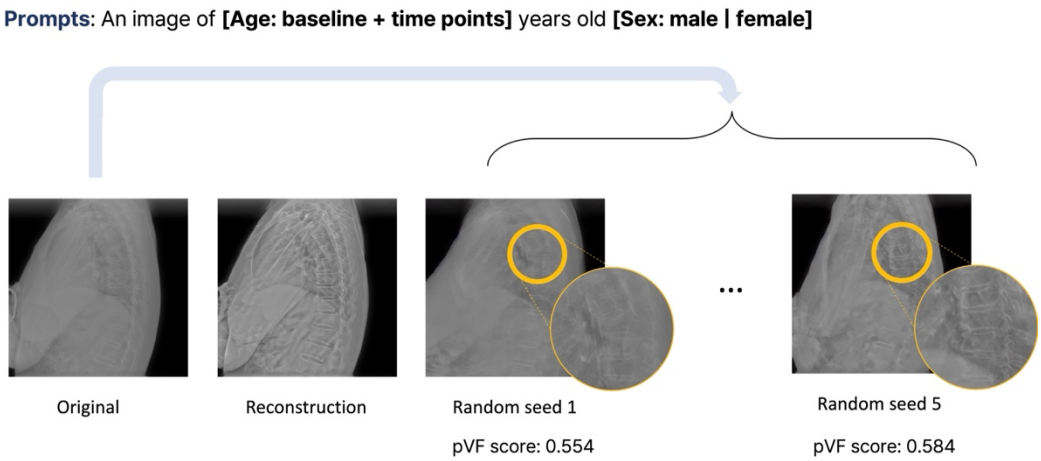


Figure 7 Generative images by age-guided prompt (when incident VF occurred) with different random seeds

2) Optimal thresholds for risk stratification

To reflect the long-term follow-up risk and ensure a fair comparison with the FRAX score, the generative pVF score was set as 10-year fracture score. Risk groups were divided into a high-risk score group and a low-risk score group for BpVF and GpVF. To stratify BpVF, the dichotomized classification threshold to determine sensitivity and specificity was set at the default of 0.5 [36] which is the chi-squared maximized threshold for log-rank testing. [57, 58] For GpVF, we

연세대학교
YONSEI UNIVERSITY

chose a cut-off point of 0.6 based on the average increment between baseline and generative pVF scores as well as the results of a log-rank test. [57, 58]

3) Comparison between risk groups

With the optimal threshold points, participants could be divided into three combinations of risk groups using the following criteria: 1) low risk for both BpVF & GpVF (LL risk group), 2) low risk for BpVF & high risk for GpVF (LH risk group), 3) and high risk for BpVF regardless of GpVF status(HIGH risk group). The findings informed this decision of study by Mills ES et al.[59], which reported that 21.9% of patients who did not receive anti-osteoporotic medications experienced secondary fractures after a vertebral osteoporotic compression fracture. This indicates that the likelihood of a subsequent fracture is significantly high after an initial fracture event. Therefore, high BpVF scores were considered high risk, regardless of their GpVF scores, to reflect the increased likelihood of fractures in the future

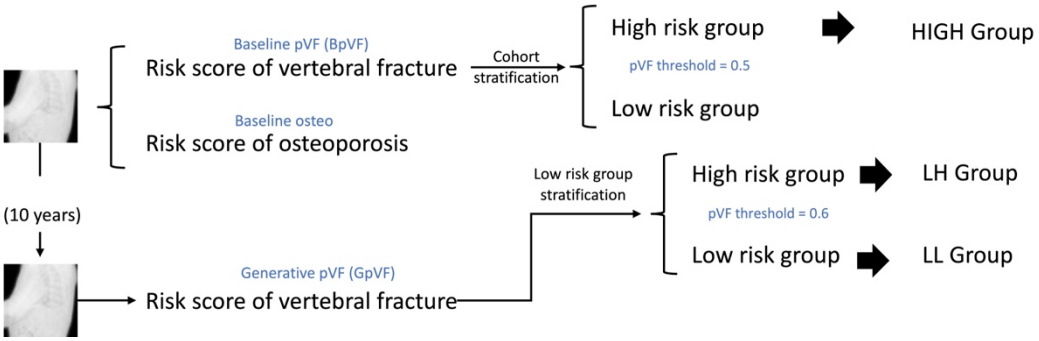


Figure 8 Risk stratification

연세대학교
YONSEI UNIVERSITY

8. Statistical analysis

In the participants' characteristics, continuous variables were presented as mean ± standard deviation, while categorical variables were shown as counts and proportions. Independent two-sample t-tests were used for comparing continuous variables, and chi-square tests for categorical variables in clinical characteristics. The areas under the receiver-operating-characteristic curves (AUROCs) for incident fracture risk prediction were compared using the DeLong method. [60] To evaluate the accuracy of fracture event predictions using the optimal threshold, metrics such as accuracy, sensitivity, specificity, and precision were calculated using confusion matrices. Statistical significance was set at a two-sided p-value of <0.05.

For the evaluation of Diff-X image generation, three statistical methods were utilized. First, the Multi-Scale Structure Similarity Index Measure (MS-SSIM) [61] and Peak Signal-to-Noise Ratio (PSNR) were used to evaluate the similarity and quality of the reconstructed image from the first LDM, respectively. Second, the accuracy of predicted age was evaluated following a previous study [44], using a regression model backboned with Efficient-net [51] to predict the age of images generated by the second LDM. Finally, a comparison between real follow-up X-ray images and generative images at the last follow-up date involved a random sample of 10% of participants from the test set. The two scores were plotted in a 2D space, and a Bland-Altman analysis was performed.

For the time-dependent survival analysis, Cox proportional hazard (CoxPH) model and the DeepSurv model [62] were employed to examine the relationship between incident fracture events and risk groups. The initial model was adjusted for three risk groups as categorical variables, and then clinical variables, age, sex,

연세대학교
YONSEI UNIVERSITY

and BMI (Model1), and FRAX major osteoporotic scores (Model2) were added to the Cox proportional hazards model, respectively. The performance of these models was evaluated by the concordance index (c-index) [63] and time-dependent AUROC with inverse probability censoring weighting (IPCW) [64]. Kaplan-Meier curves were plotted for the risk group for incident fractures. Statistical analyses were conducted using the Scipy 1.8 library [65] in Python (version 3.8) and R (version 3.6.3).

연세대학교
YONSEI UNIVERSITY

III. RESULTS

1. Clinical characteristics of the study subjects

In this study, a total of 9,726 individuals were involved in the final analysis. The mean age of study participants in the derived cohort was 65.7 years, with females comprising 66% (6,105 individuals) of the cohort. Morphometric vertebral fractures were observed in 18.6% of the participants, while 10.3% experienced incident fractures during the follow-up period. Due to stratified random sampling, no significant differences were observed between the training, validation, and test sets ($p > 0.05$).

Lateral spine X-ray images were sourced from various manufacturers, predominantly General Electric (GE), which contributed to 70% (18,738 images) of the total. To enhance the generalizability of our model, these X-ray images included not just the lumbar spine but also the thoracic, sacrum, and cervical spine areas. The combination of thoracic and lumbar spine images made up 68% of the total dataset. Further details can be found in the table listing the manufacturers and the visualized areas of X-ray images.

Table 1 Clinical characteristics of study participants

	Derivation of cohort			
	Overall (n=9276)	Train set (n=5568, 60%)	Validation set (n=1856, 20%)	Test set (n=1852, 20%)
Women, n (%)	6105 (66)	1219 (66)	1220 (66)	3666 (66)
Age, years	65.7 ± 8.5	65.7 ± 8.5	65.8 ± 8.5	65.7 ± 8.5
Height, cm	159.1 ± 8.5	159.2 ± 8.5	159.2 ± 8.4	158.8 ± 8.6
Weight, kg	61.1 ± 10.6	61.1 ± 10.6	61.2 ± 10.6	61.0 ± 10.3
BMI, kg/m2	24.0 ± 3.2	24.0 ± 3.2	24.1 ± 3.3	24.1 ± 3.2
Lumbar spine BMD (g/cm2)*	0.895 ± 0.212	0.894 ±0.215	0.894 ± 0.209	0.899 ± 0.221
Lumbar score T-score*†	-1.4 ± 1.9	-1.4 ±1.9	-1.3 ± 2.0	-1.4 ± 2.0

연세대학교
YONSEI UNIVERSITY

Femoral neck BMD (g/cm2)*	0.654 ± 0.138	0.650 ± 0.138	0.654 ± 0.137	0.657 ± 0.141
Femoral neck T-score*†	-1.7 ± 1.2	-1.7 ± 1.1	-1.7 ± 1.2	-1.7 ± 1.2
Total hip BMD (g/cm2)*	0.792 ± 0.155	0.788 ± 0.153	0.792 ± 0.154	0.795 ± 0.159
Total hip T-score*†	-1.2 ± 1.3	-1.2 ± 1.3	-1.2 ± 1.3	-1.2 ± 1.3
Osteoporosis, n (%)*	2649 (40.3)	1590 (40.3)	533 (40.1)	526 (40.1)
Morphometric vertebral fracture, n (%)	1723 (18.6)	1035 (18.6)	349 (18.8)	339 (18.3)
Any incident fracture, n (%)	921 (10.3)	546 (10.3)	184 (10.3)	191 (10.3)
Vertebral fracture	705 (7.6)	426 (7.6)	139 (7.6)	140 (7.6)
Non-vertebral fracture	291 (3.6)	159 (3.6)	65 (3.6)	67 (3.6)
Follow up duration, days (median)	766	772	750	765
History of fracture (clinical), n (%)	682 (7.4)	426 (7.7)	124 (6.7)	132 (7.1)
Glucocorticoid users, n (%)	437 (4.7)	250 (4.5)	86 (4.6)	101 (5.5)
Rheumatoid arthritis, n (%)	252 (2.7)	145 (2.6)	46 (2.5)	61 (3.3)
Secondary osteoporosis**	188 (2.0)	119 (2.1)	31 (1.7)	38 (2.1)

Table 2 Manufacturer types and visualized area of spine x-ray images in derivation

	Derivation cohort, image n=26,299
General Electric (GE)	18,738 (71)
DongKang (DK)	2,628 (10)
Fuji electric	2,024 (8)
Samsung Electronics	1,371 (5)
KODAK	1,274 (5)
Listem	110 (<1)
Toshiba	73 (<1)
JSB	48 (<1)
Carestream	14 (<1)

연세대학교
YONSEI UNIVERSITY

ADT	7 (<1)
Canon Inc	6 (<1)
SIEMENS	6 (<1)
T-L	7,482 (28)
T-L-S	4,130 (16)
L-S	6,281 (24)
Whole spine	6,825 (25)
C-T-L	1,528 (6)
T	44 (<1)
C-T	9 (<1)

2. Follow-up X-ray Image generation

Diff-X Sampling

The Diff-X model was utilized to generate future follow-up images based on baseline images and corresponding clinical prompts at predetermined intervals (1 year, 5 years, 10 years, and the last follow-up date). Figure 10 illustrates the overall results of images produced by Diff-X, compared to a single text-based Latent Diffusion Model (LDM) [28]. Given that our model builds upon a text-based diffusion model, the single LDM without feature injection serves as a baseline for comparison.

In our observations, images generated by Diff-X exhibited superior quality compared to those from the single LDM. The single LDM tended to produce partially blurry images, often failing to generate either accurate reconstructions or discernible age-related changes. In contrast, Diff-X effectively demonstrated variations in the spine across different ages. A notable comparison is the portrayal of fractures by both models. Interestingly, both models identified potential fractures at the same locations. However, the single LDM distorted the spine's morphological structure and produced ambiguous representations of fractures whereas Diff-X maintained greater fidelity to the input image's structure.

연세대학교
YONSEI UNIVERSITY

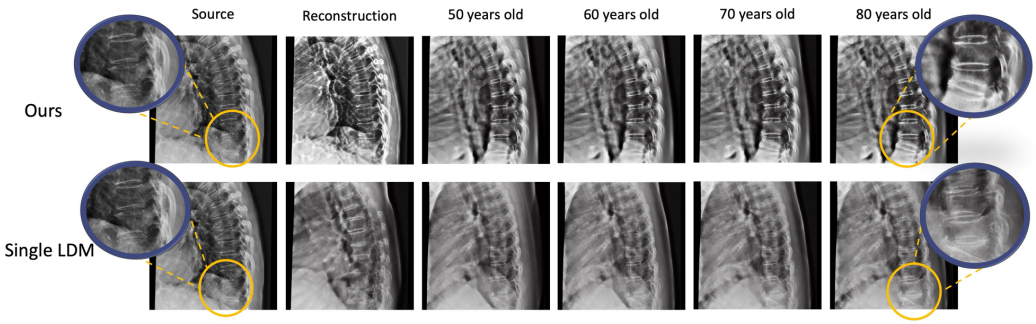


Figure 9 Source images of a 79 years old male and generated samples of spine X-ray images using our model and a single LDM

Performances of Reconstruction and Age Prediction

The reconstruction quality was quantitatively assessed by comparing the Multi-Scale Structure Similarity Index Measure (MS-SSIM) and Peak Signal-to-Noise Ratio (PSNR) between the input image and the reconstruction image from the first Latent Diffusion Model (LDM). Diff-X achieved an MS-SSIM score of 0.506 and a PSNR of 20.737, outperforming the single LDM, which scored 0.219 in MS-SSIM and 15.856 in PSNR. This indicates that Diff-X had superior performance in both MS-SSIM and PSNR compared to the single LDM.

The age prediction accuracy of the model was evaluated using deep learning, specifically an Efficient-net based approach. The correlation between actual ages and predicted ages by Efficient-net yielded an R-square of 0.841, demonstrating high accuracy in age prediction.

연세대학교
YONSEI UNIVERSITY

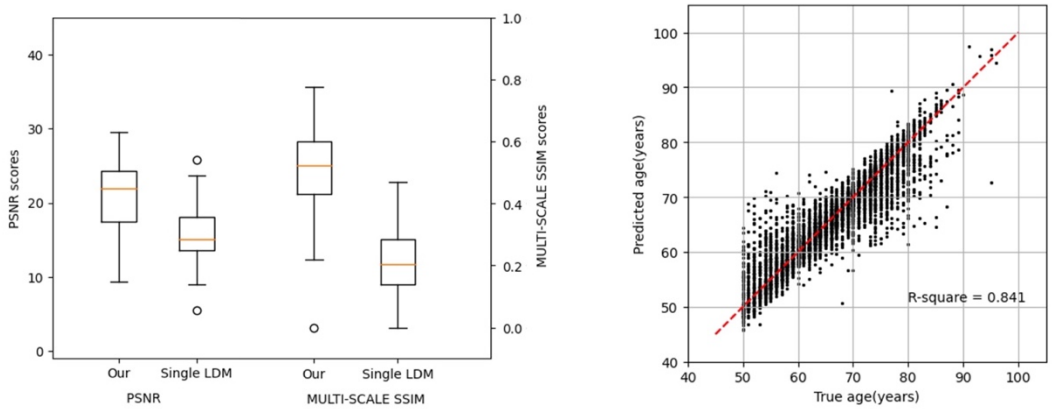


Figure 10 . Model performance evaluation. (a) Box plot of PSNR and Multi-Scale SSIM score for the image reconstruction (b) Scatter plot of regression results for X-ray images

In Figure 10, (b) illustrates the differences between two generated images, with red highlighting changes associated with older age. The age-translation images show noticeable differences, particularly in the spine. Typically, images depicting older age exhibit features like compression fractures or calcification.

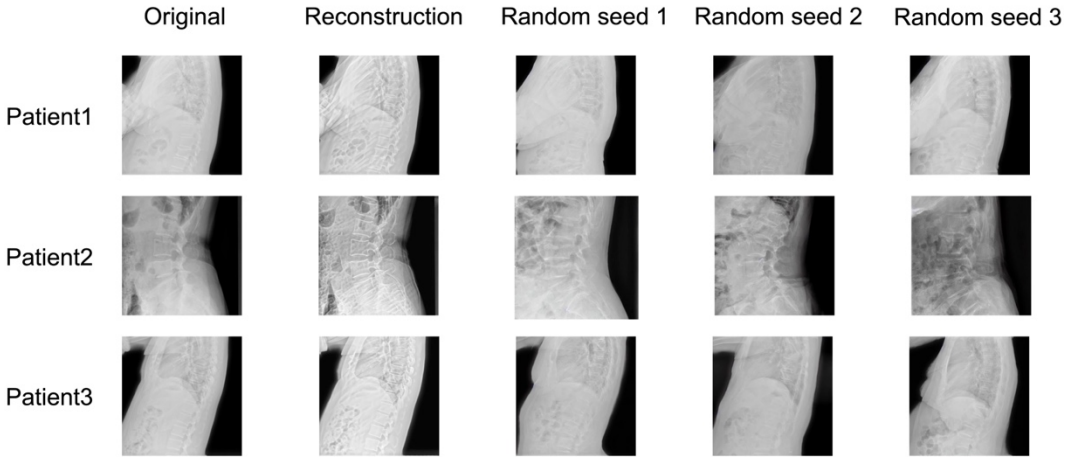


Figure 11 Generative future-possible X-ray images by age guided prompts with different random seeds.

연세대학교
YONSEI UNIVERSITY

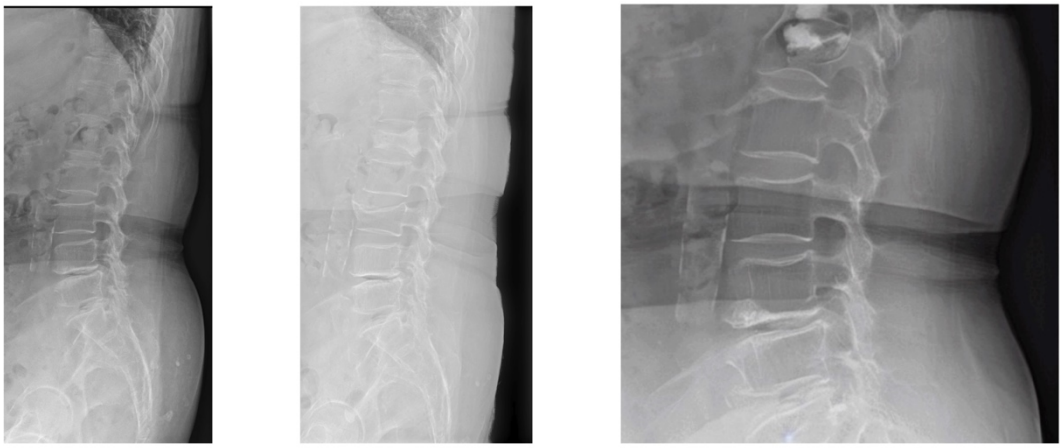


Figure 12 Generated image comparison with the real follow-up image. generative images from a baseline image using our model with the specific age point. (A) Baseline image of patient 71 years old (B) Follow-up image of the patient after 4 years (75 years old) (C) Generative images from the baseline image using our model

Feature Extraction

To demonstrate the manipulation of latent spaces, feature maps from each layer at different timesteps were extracted, as depicted in Figure 13. Principal Component Analysis (PCA) was applied to these extracted features, and the first three principal components were visualized as RGB channels. To maintain consistency with X-ray images, these visualizations were converted to grayscale. This approach was instrumental in understanding how spatial information was preserved in the first LDM and determining the optimal layers for integration into the second LDM.

In self-attention layer 5 and Resnet layer 5, the shape of vertebral bodies in the spine area began to manifest, and the surrounding soft tissues became more distinct. Furthermore, in self-attention layer 8 and Resnet layer 8, the edges of each individual vertebral body were more sharply defined and easier to identify.

연세대학교
YONSEI UNIVERSITY

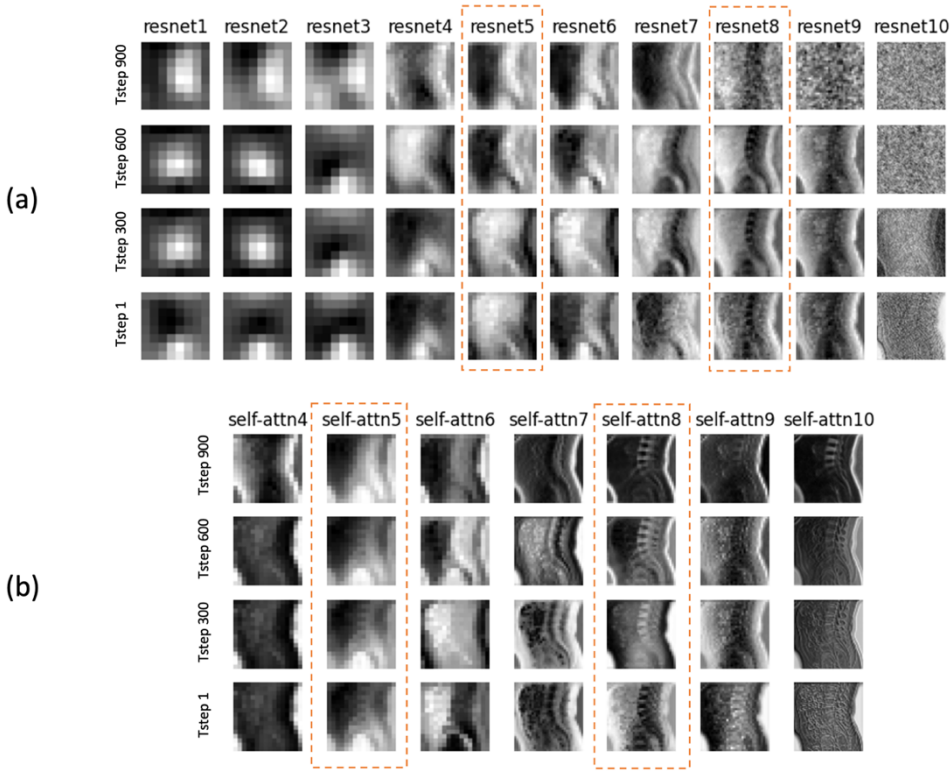


Figure 13 Extracted 3 top components using PCA from the resnet blocks and self-attention modules were visualized in the feature maps (a) 10 Resnet output features (b) Self-attention block output features from the first LDM in diffusion timesteps 900, 600, 300, and 1

3. Vertex pVF score of baseline and generative images

Among the 1,852 participants in the test set, there were 191 fracture cases reported over a median follow-up period of 765 days (Interquartile Range, IQR: 1142.5 days). The mean BpVF and GpVF over the 10-year period were 0.491 and 0.600 for people with new fractures and 0.39 and 0.52 for people without new fractures.

연세대학교
YONSEI UNIVERSITY

A regression analysis exploring the linear relationship between the pVF scores of real and generative follow-up X-ray images yielded a correlation coefficient of 0.656 in a dataset sampled from 10% of the test set. Furthermore, a Bland-Altman analysis revealed a mean difference of -0.06 and a standard deviation of 0.2, indicating a high level of concordance between the actual pVF scores and the generative pVF scores.

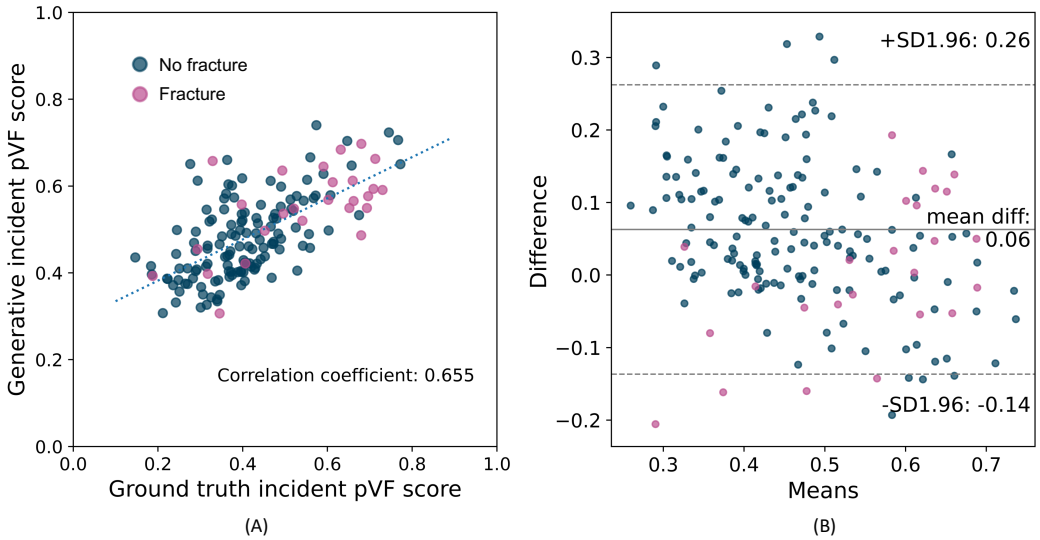


Figure 14 Scores comparison between real pVF score and generative pVF score at the same time point. (A) Plot of linear regression analysis which indicates a linear relationship between the pVF score of real follow-up and generative X-ray images with correlation coefficient = 0.656 in the sampling dataset. (B) The Bland and Altman plot showing the relationship between the difference and mean of the pVF score of real follow-up and generative X-ray images.

While there is no strong correlation between BpVF and GpVF at 10 years (correlation coefficient = 0.376), a high correlation was observed between GpVF at different time points (correlation coefficient 1 year vs 5 years 0.735; 1 year vs 10 years 0.694; 5 years vs 10 years 0.721). Additionally, age was found to have a high correlation with GpVF

연세대학교
YONSEI UNIVERSITY

(correlation coefficient = 0.668). From baseline time points to 10 years, the scores increased by an average of 0.103 points.

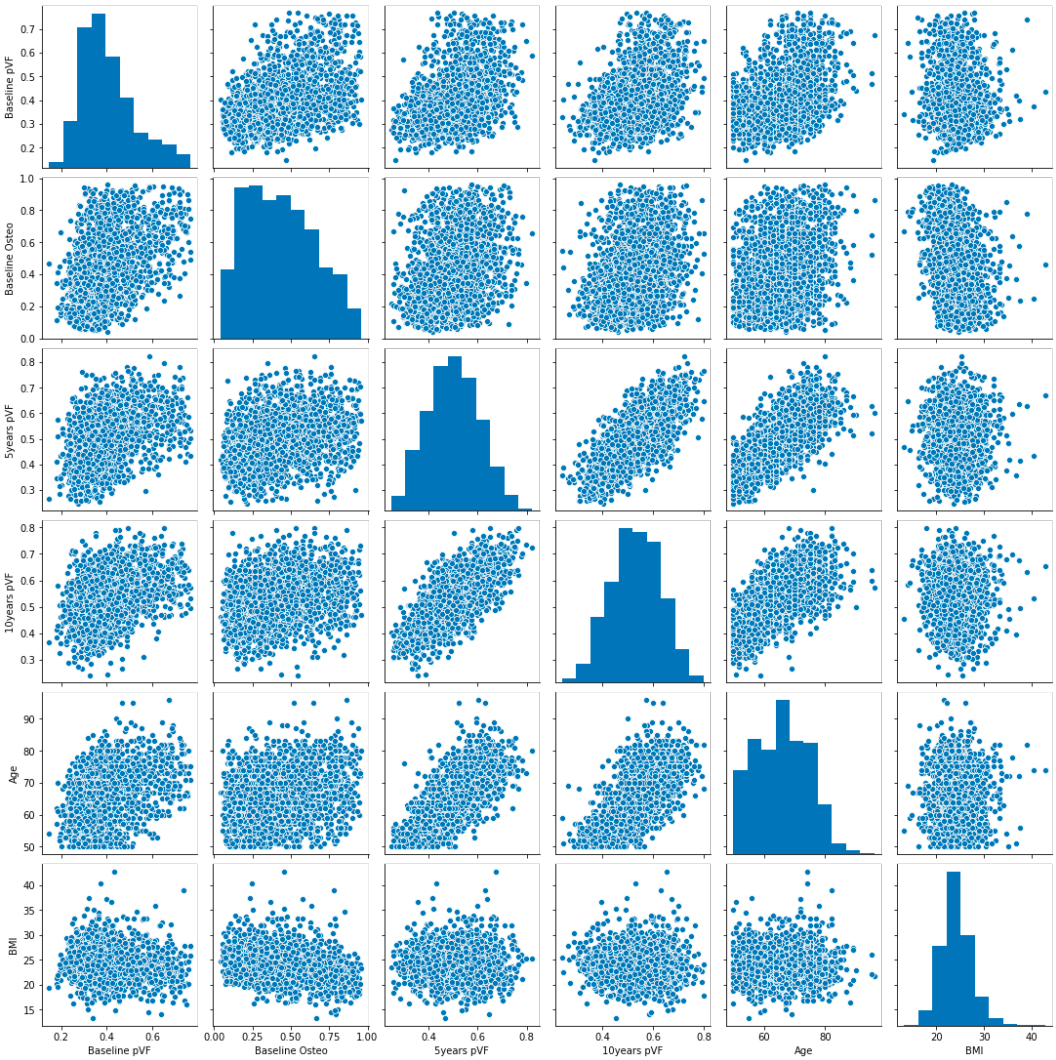


Figure 15 Pair plot to find the relationship between the baseline pVF score and the generative scores

연세대학교
YONSEI UNIVERSITY

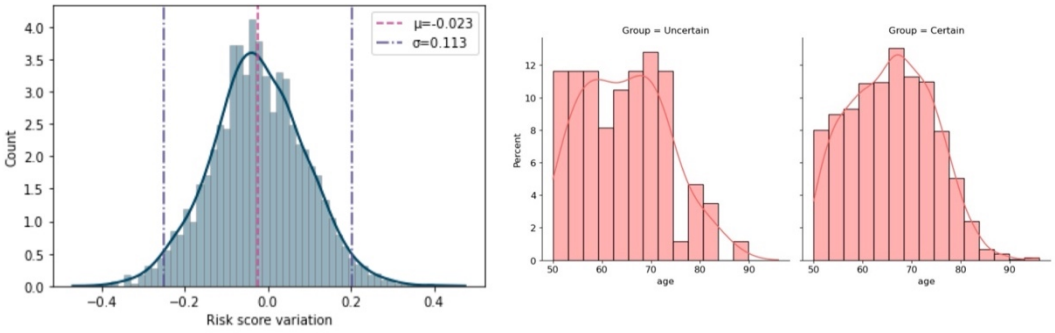


Figure 16 The uncertainty of generative pVF score. (a) Distribution of generative fracture score variation (b) Age comparison between uncertain group and certain group

When calculating the variation of GpVF and plotting its distribution, results similar to those shown in Figure 16 can be obtained, which closely resemble a normal distribution. When considering individuals beyond two standard deviations as out of distribution, or highly uncertain, it was observed that this uncertain group tended to be younger compared to the certain group, with a higher proportion of patients in their 50s to early 60s.

4. Incident fracture prediction model

The DeepSurv survival models built on BpVF and GpVF (Fracture scores model) demonstrated comparable performance in detecting incident fractures over a 10-year period when compared to the FRAX score (Mean time-dependent AUROC, 0.69 and 0.69, respectively). The Fracture scores model exhibited superior performance to the FRAX model before 5 years, while the FRAX model performed better in the remaining follow-up period (time-dependent AUROC at 1 year: Fracture scores model 0.76, FRAX model 0.70; at 5 years: Fracture scores model 0.70, FRAX model 0.72; at 10 years: Fracture scores model 0.62, FRAX model 0.70). When age, sex, and BMI were added to the Fracture scores model, its discriminatory ability improved, surpassing both the Fracture scores and FRAX model (Mean time-dependent AUROC 0.74; time-dependent AUROC

연세대학교
YONSEI UNIVERSITY

at 1 year 0.79; at 5 years 0.76; at 10 years 0.74). These improvements remained robust over the entire 10-year period.

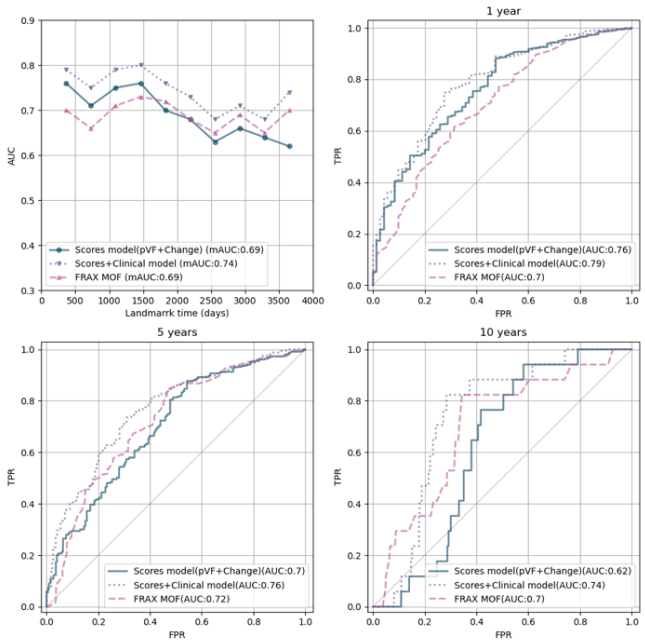


Figure 17 Comparison of the time-dependent area under the curve (AUC) and receiver operating characteristic (ROC) curves of models by Deep learning-based Cox models consisted of 1) fracture scores 2) 1) with age, sex, and BMI, and 3) FRAX major osteoporotic at 10 years. (A) Time-dependent AUC and (B, C, D) ROC curves at 1 year, 5 years, and 10 years. Mean AUC values are shown in the legend.

The fracture scores model and clinical added model showed similar trajectories in the time-dependent AUROC for CoxPH and demonstrated lower performance compared to DeepSurv-based survival models.

연세대학교
YONSEI UNIVERSITY

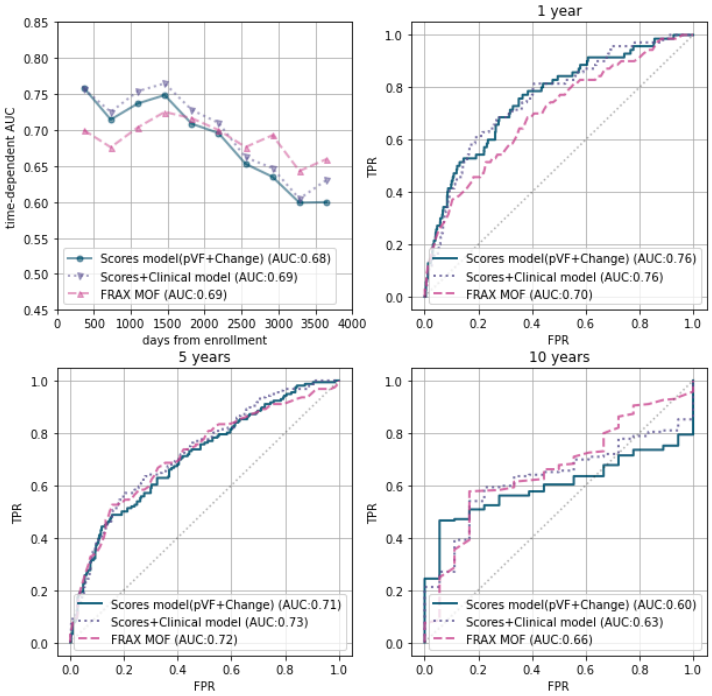


Figure 18 Comparison of the time-dependent area under the curve (AUC) and receiver operating characteristic (ROC) curves of models by Cox proportional hazard function consisting of 1) fracture scores 2) 1) with age, sex and BMI, and 3) FRAX major osteoporotic at 10 years. (A) Time-dependent AUC and (B, C, D) ROC curve at 1 year, 5 years and 10 years. Mean AUC values are shown in the legend.

5. Risk stratification for fracture risk

Based on the cut-off values for BpVF and GpVF scores, participants were divided into three risk groups: Low-Low (LL), Low-High (LH), and High (HIGH). The distribution of individuals in these groups was 1182 (63.8%), 322 (17.4%), and 348 (18.8%), respectively. The 10-year fracture incidence rates for each group were 5.9% for LL, 11.2% for LH, and 24.4% for HIGH, indicating higher rates in groups with elevated BpVF scores. The LL group, which had the lowest BpVF scores, demonstrated a distinction in fracture rates when further stratified by

연세대학교
YONSEI UNIVERSITY

GpVF scores. Furthermore, the LL group was observed to be the youngest, with the highest average height and weight (160.3 cm and 62.2 kg), compared to the LH (156.5 cm and 59.7 kg) and HIGH (154 cm and 57.0 kg) groups ($p < 0.001$). Among those groups, bone mineral density showed a significant difference. Distinct clinical characteristics among these groups are detailed in Table 3.

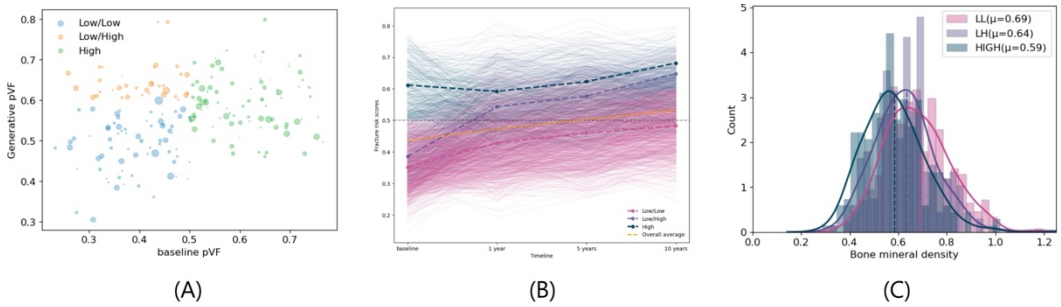


Figure 19 (A) Fracture scores distribution in our test set. (B) Time-dependent trajectory of image-based fracture scores in different risk groups (C) Bone mineral density in different risk groups

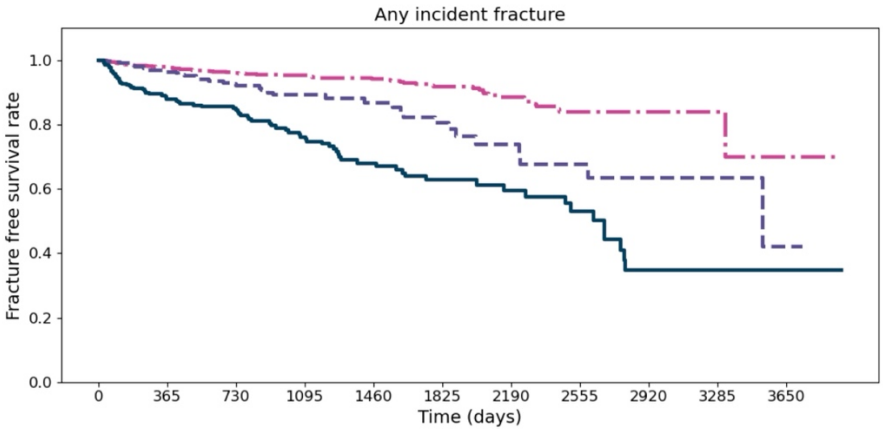
Table 3 Clinical characteristics of different risk groups

	Low/Low	Low/High	High
Group n	1182(63.8)	322(17.4)	348(18.8)
Prevalent fracture (%)	89(7.5)	41(12.7)	267(76.7)
Incident fracture (%)	70(5.9)	36(11.2)	85(24.4)
Woman (%)	693(58.6)	251(78.0)	276(79.3)
Age	63 ± 7.6	71 ± 5.8	71 ± 7.7
Height	160.3 ± 8.5	156.5 ± 7.3	155.6 ± 8.8
Weight	62.2 ± 10.3	59.7 ± 9.5	58.0 ± 10.3
BMI	24.1 ± 3.0	24.4 ± 3.5	23.9 ± 3.6
FRAX MOF	0.06 ± 0.04	0.08 ± 0.05	0.1 ± 0.06
FRAX HF	0.02 ± 0.02	0.03 ± 0.03	0.05 ± 0.04
FNBMD	0.69 ± 0.18	0.64 ± 0.15	0.59 ± 0.14

연세대학교
YONSEI UNIVERSITY

Previous fracture	30(2.5)	16(5.0)	86(24.7)
Glucocorticoid	66(5.6)	11(3.4)	24(6.9)
Rheumatoid arthritis	34(2.9)	13(4.0)	14(4.0)
Secondary osteoporosis	22(1.9)	5(1.6)	11(3.2)

Kaplan-Meier curves and the log-rank test showed significant p-values for risk group comparisons ($p < 0.001$ for all other group comparisons).



At risk/Years	0	1	2	3	4	5	6	7	8	9	10
LL	1173	944	693	467	350	223	145	81	29	7	1
LH	318	245	169	103	68	42	25	17	10	6	2
HIGH	343	220	162	106	69	53	34	22	11	5	1

Figure 20 Kaplan Meier plot for risk stratification using pVF scores at baseline and after 10 years. Categorization of three risk groups using the following criteria: Low-Low (LL) for individuals with both low BpVF and GpVF scores, Low-High (LH) for those with low BpVF scores but high GpVF scores, and HIGH for individuals with high BpVF scores. These risk groups were established by applying cutoff points of 0.5 for BpVF and 0.6 for GpVF. Notably, the log-rank test yielded significant p-values for group comparisons: $p < 0.001$ for all other group comparisons

연세대학교
YONSEI UNIVERSITY

A Cox regression analysis revealed that, compared to the LL group, individuals in the LH and HIGH groups had 109% and 391% increased risks of fracture, respectively (hazard ratios [HR] of 2.090 and 4.911; P=<0.001 for all). These risk groups remained significant predictors of fracture risk when clinical variables (age, sex, and BMI) or FRAX major osteoporotic scores were included (HR of 1.598 and 3.726 for clinical variables, and 1.643 and 3.235 for FRAX, respectively; P=<0.05 for all). The chi-square for the likelihood ratio in the Cox model improved when stratifying risk into three groups compared to using only BpVF with a 0.5 threshold (Likelihood ratio Chi-square 81.07 vs 97.9; p<0.001).

Table 4 Univariate and multivariable analysis for Cox proportional hazard regression models on the predictors of incident fracture in the clinical test set

Predictors of incident fracture	Univariable model		Multivariable model 1		Multivariable model 2	
	Hazard Ratio (95% CI)	P value	Hazard Ratio (95% CI)	P value	Hazard Ratio (95% CI)	P value
Risk group						
LL	1.000 (Ref)		1.000 (Ref)		1.000 (Ref)	
LH	2.090(1.396,3.131)	<0.001	1.598(1.044,2.448)	0.031	1.643(1.090,2.476)	0.018
HIGH	4.911(3.574,6.750)	<0.001	3.726(2.622,5.296)	<0.001	3.235(2.289,4.572)	<0.001

연세대학교
YONSEI UNIVERSITY

Age (per 1-year increase)	1.064(1.046,1.082)	<0.001	1.036(1.016,1.056)	0.002		
SEX (female versus male)	1.846(1.283,2.657)	<0.001	1.466(1.014,2.121)	0.063		
BMI (per 1kg/m2 increase)	0.986(0.943,1.031)	0.542	0.987(0.946,1.029)	0.443		
FRAX major osteoporotic (per 1 SD increase)	1.663(1.528,1.81)	<0.001			1.461(1.324,1.612)	<0.001
C-index	Clinical (Age,Sex,BMI): 0.654 FRAX: 0.690 Risk group alone: 0.675 Model1: 0.700 Model2: 0.727					
Likelihood ratio chi-square	Clinical (Age,Sex,BMI): 61.0 FRAX: 93.58 Risk group alone: 92.7 Model1: 105.7 Model2: 137.0					

Inclusion of BpVF and GpVF scores as continuous variables in the Cox model showed that a 1 SD increase in fracture scores (0.1 point) was significantly

연세대학교
YONSEI UNIVERSITY

associated with increased fracture risk (HR of 1.824 with 95% CI 1.601–2.079 for BpVF; 1.284 with 95% CI 1.093–1.509 for GpVF; $p < 0.005$).

Furthermore, GRAD-CAM analysis indicated that the DCNNs were primarily focused on pixel values from vertebral bone regions. Notably, in cases with high GpVF scores, GRAD-CAM emphasized specific fractured regions as the most prominently changed areas.

Table 5 Univariate and multivariable analysis for Cox proportional hazard regression models on the predictors of incident fracture in the clinical test set

Predictors of incident fracture	Univariable model		Multivariable score-based model		Multivariable clinical model	
	Unadjusted HR (95% CI)	P value	Adjusted HR (95% CI)	P value	Adjusted HR (95% CI)	P value
Baseline pVF (per 0.1-point increase)	1.945(1.721,2.199)	<0.001	1.824(1.601,2.079)	<0.001		
Generative pVF (per 0.1-point increase)	1.583(1.362,1.839)	<0.001	1.284(1.093,1.509)	0.002		
Baseline pVF Risk group (Low versus High)	4.033(3.029, 5.369)	<0.001				
Generative pVF Risk group (Low versus High)	2.304(1.721, 3.085)	<0.001				

연세대학교
YONSEI UNIVERSITY

Age (per 1-year increase)	1.064(1.046,1.082)	<0.001		1.063(1.044,1.081)	<0.001
SEX (female versus male)	1.846(1.283,2.657)	<0.001		1.748(1.214,2.516)	0.003
BMI (per 1kg/m2 increase)	0.986(0.943,1.031)	0.542		0.987(0.945,1.030)	0.545
C-index	Baseline pVF Risk group: 0.648 Score-based model: 0.726 Clinical model: 0.654				
Likelihood ratio chi-square	Baseline pVF Risk group: 81.07 Score-based model: 111.2 Clinical model: 60.96				

연세대학교
YONSEI UNIVERSITY

IV. DISCUSSION

In this study, we developed a language-guided image generative model, Diff-X, to create future X-ray images and calculate baseline (BpVF) and generative (GpVF) fracture scores in a substantial hospital-based cohort. The survival model we employed successfully predicted incident fractures over a 10-year period. Furthermore, our analysis demonstrated the effectiveness of risk group stratification based on these scores. We found that dividing the cohort into three risk groups (LL, LH, and HIGH) significantly improved the likelihood ratio chi-square values compared to binary stratification using only BpVF, alongside demographic information. Notably, each of these three risk groups maintained its predictive independence even when the FRAX score was incorporated into the Cox model.

This study underscores the potential of generative models for predicting fracture risk from X-ray images and visualizing the possible evolution of spinal conditions. The use of generative models like Diff-X offers a novel approach to risk stratification, potentially enhancing the accuracy and utility of clinical decision-making in the context of osteoporotic fracture risk.

The generative model, "Diff-X", was developed by adapting the "Play and Plug" model [47] which is a framework for text-to-image synthesis in the context of image-to-image translation. We modified this model to use low-resolution images and different positions of feature insertion. The original model [47] highlighted challenges with texture-less images and latent encoding of dominant low-frequency appearance information in the DDIM scheduler. In our experiments, we observed that without preserving the original semantic layout, the details in X-ray images, such as spine structure, soft tissues, and vertebrae count, were easily

연세대학교
YONSEI UNIVERSITY

altered by text prompts and random seeds. To address this, we incorporated a low-resolution version of the original image into the first LDM, thereby improving the reconstruction quality of baseline X-ray images. Furthermore, the previous study[47] mainly focused on RGB three-channel images, such as human figures, buildings, and scenes. Our study investigated adapting this framework to one-channel grayscale medical images. Through modifications, we demonstrated that our text-guided image generative model could produce potential future images based on baseline images.

Few studies have focused on predicting temporal changes in medical images due to uncertainty. While GANs [66] and mathematical models [67] have been used to synthesize realistic data with time changes, they often pose challenges in training and clinical applicability. Our diffusion-based generative model approach addressed these issues. Pinaya [44] used diffusion models to generate different brain T1 MRI images and structural volumes guided by numerical data (coefficient 0.692). Our research also demonstrates the potential of diffusion-based generative models for fracture risk stratification and guiding individualized therapies for fracture prevention. The applicability of these models in medical imaging is an active area of research, as seen in the proceedings of MICCAI 2023 and advancements in Text-to-image generation techniques by OPEN AI's "DALLE 3" [68] and ChatGPT [69, 70]. Further studies are needed to validate their clinical utility.

For input baseline X-ray images, we included extra-spinal regions to train our model. While spinal fractures are primary considerations, soft tissues and muscles are also critical for predicting fracture risk, as osteoporosis has strong correlations with fat and muscle. [71-74] In unpublished experiments, we observed X-ray

연세대학교
YONSEI UNIVERSITY

images' potential to predict body composition parameters from CT images. As indicated by SHAP values [75], skeletal muscle density along with the density and area of subcutaneous fat, are indicators of osteoporosis risk. Therefore, using entire X-ray images was more appropriate than focusing solely on the spine.

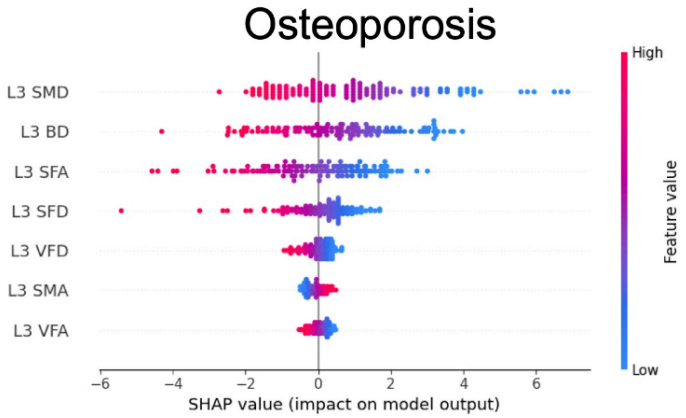


Figure 21 SHapley Additive exPlanations (SHAP) value of prediction models using CT body composition parameters to predict the presence of osteoporosis. BMI, body mass index; SMA, skeletal muscle area; SMD, skeletal muscle density; VFA, visceral fat area; VFD, visceral fat density; SFA, subcutaneous fat area; SFD, subcutaneous fat density; BD, bone density.

Table 6 Prediction performance of body composition using X-ray images

	High visceral fat	Low muscle mass	Low muscle desity	Low bone density	Low subcutaneous fat
CXR-Clinical	0.921 (0.911,0.929)	0.799 (0.783,0.816)	0.862 (0.848,0.875)	0.863 (0.849,0.877)	0.923 (0.912,0.933)

연세대학교
YONSEI UNIVERSITY

Deep learning models in musculoskeletal research need to be clinically applicable and provide added value. [35] Assessing the quality of generative models, including GANs and diffusion models, typically involves human evaluation. [76-78] For our model's reliability, we employed the VERTE-X pVF model to compare real follow-up and generative images objectively, without human intervention.

Based on the uncertainty calculated by GpVF, young age individuals showed a wider range of potential developments in the fucture, with scores more broadly distributed, whereas the certain group shows more consistenct results. Further research is needed to uncover the implication of the overall data distribution on the study results.

Our baseline and generative pVF scores can be computed in a minute, including the generation of future potential images. In South Korea, the General Health Screening Program (GHSP) [79] covers a comprehensive assessment, including X-rays and clinical variables like age, sex, and BMI. Our systemized model could leverage these data to establish an efficient warning system for fracture risks.

Interpretability is crucial in data-driven models, particularly in healthcare. [80] In our study, X-ray images were generated to visually explain potential future developments and the severity of spine conditions. Additionally, the pVF scores from baseline and generative images effectively stratify fracture risk groups, enhancing the clinical utility of these models.

Communicating risks to patients is critical aspect of ensuring informed consent in clinical practice. [81] Clinicians often face challenges when patients' make choices contrary to their long-term health goals, influenced by various factors like financial constraints or personal habits. [82] Fractures, as a long-term health

연세대학교
YONSEI UNIVERSITY

concern, require specific strategies to improve patient understanding of associated risks. Utilizing positive and negative case outcomes and appropriate visual aids is recommended. [81, 83, 84] Our generative images could serve as such aids, providing tangible visualizations of a patient's future health status. By aligning these images with a patient's specific risk group, clinicians can offer tailored therapies for fracture prevention and communicate risks more effectively.

Our model demonstrated superior performance in predicting incident fractures compared to previous studies. [31, 32] Hsieh's research, which used deep learning to calculate hip and spine BMD scores, did not account for time effects on incident fracture prediction and was limited to lumbar spine images. Kong's research, while predicting incident fractures using lumbar spine radiographs (c-index: 0.612), did not consider other spine areas such as thoracic, sacrum, and cervical regions. Our study's risk stratification Cox model achieved a c-index of 0.679, which improved to 0.729 with the addition of clinical variables. The likelihood ratio chi-square was also higher in our models. Although the fusion of image data and clinical metadata has been effective in other contexts[85], it did not significantly improve performance compared to image-only models for some tasks.[86] Our findings suggest that while imaging data are crucial for detecting prevalent vertebral fractures and osteoporosis, clinical features may provide only minimal additional benefit. However, integrating imaging data with clinical metadata might improve the model's performance in predicting incident fracture risk, as clinical risk factors have been shown to enhance fracture risk prediction when added to bone density. [87]

Kaplan-Meier plots and hazard ratios from the Cox model indicated that the highest risk group was HIGH, followed by LH, and LL, with significant

연세대학교
YONSEI UNIVERSITY

differences. The higher-risk groups were older, had lower BMD, and had higher FRAX scores, consistent with clinical expectations. [88] This suggests that our generative model can create age-related images aligned with clinical risk factors. Our results can be integrated into clinical practice to stratify patients at high risk of incident fractures. While BMD is a key fracture predictor, DXA scans can sometimes be unreliable.[89] Our findings suggest that X-ray images, a common diagnostic tool, could help identify future fracture risks.

In evaluating the cost-benefits for incident fracture detection using BpVF and GpVF scores, it becomes evident how much the current FRAX-based fracture risk assessment process can be improved. Generally, in fracture risk measurement, a FRAX Major Osteoporotic Score above 20% indicates a very high risk of fracture. An examination of this in a test set of 1,852 individuals classified 53 as high-risk and 1,799 as low-risk for fractures. Of these, 21 fractures occurred in the high-risk group, while 170 fractures occurred in the low-risk group. When projected onto a group of 100 individuals, although 10 people experienced fractures, appropriate treatment was administered in only 1 case, leaving 9 untreated. However, additional classification of the 1,799 untreated individuals using BpVF and GpVF scores could potentially enable treatment for another 635 individuals. Of these, 105 actually experienced fractures, facilitating treatment for about 70% of the total fracture group. Additionally, the treated group, including those without actual fractures, had statistically significantly lower Bone Mineral Density (BMD) than the untreated group, suggesting that these individuals could be considered at risk of fracture due to low BMD. The overall integration of GpVF into FRAX lowered the specificity for incident fracture detection due to the expanded treatment group. However, the sensitivity metric, as well as the F1 score which

연세대학교
YONSEI UNIVERSITY

considers both precision and sensitivity, showed a progressive increase. This indicates a more accurate identification of the fracture group.

Overall, our language-guided diffusion model designed to generate follow-up X-ray images of patients over aging and calculate personalized Verte-X pVF scores based on both baseline images and generative follow-up images. Based on our research, the calculated pVF scores represented a quantified assessment of fracture risk. Clinicians can use those data driven-scores to design different "personalized treatment plans" for each patient. While the current study has effectively demonstrated the use of pVF scores for stratifying risk groups, further research is required to determine which specific medications and treatments are most effective in accordance with the generated scores.

This study has several limitations that warrant consideration. Firstly, the cohort was derived from spine radiography data at tertiary-level institutions, which may limit the generalizability of our findings. External validation in primary care settings is needed to confirm the applicability of our results to a broader population. Additionally, to ensure the clinical relevance and quality of the dataset labeling process, we restricted our derived set to 26,299 X-ray scans from 9,276 individuals aged 50 and older who had follow-up spine X-rays. Consequently, data from about 40,000 individuals younger than 50 or without follow-up X-ray scans were not included. Expanding the dataset to include these individuals could potentially enhance the performance and robustness of the models.

Another limitation is that the study population was exclusively Korean, so the models should be tested on non-Korean populations to evaluate their universal applicability. Moreover, we did not grade the severity of vertebral fractures at the current stage of our research. A more detailed labeling process for fracture

연세대학교
YONSEI UNIVERSITY

severity and subsequent analysis of the correlation between VERTE-X pVF scores could provide deeper insights into the interpretation of these scores.

Finally, to simplify the preprocessing steps, we solely used lateral spine radiographs. Incorporating posterior-anterior views could unlock new possibilities and potentially strengthen our results; however, this aspect requires further investigation. [90]

연세대학교
YONSEI UNIVERSITY

V. CONCLUSION

In conclusion, our diffusion-based, language-guided image generative model successfully generated future images and calculated baseline and generative fracture scores in a large hospital-based cohort. The survival prediction model demonstrated enhanced performance in predicting incident fractures over a 10-year period. Furthermore, our analysis revealed that a three-group division significantly improved the likelihood ratio chi-square values compared to a binary stratification based solely on BpVF and demographic information.

This study underscores the potential of generative models in predicting fracture risk from X-ray images and in visualizing the potential progression of spinal conditions.

연세대학교
YONSEI UNIVERSITY

VI. REFERENCES

1. Cummings, S. R. & Melton, L. J.: Epidemiology and outcomes of osteoporotic fractures. The Lancet **359**, 1761-1767, (2002).
2. Johnell, O. & Kanis, J.: An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporosis international **17**, 1726-1733, (2006).
3. Kanis, J. A. et al.: A systematic review of hip fracture incidence and probability of fracture worldwide. Osteoporosis international **23**, 2239-2256, (2012).
4. Johnell, O. & Kanis, J. A.: An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int **17**, 1726-1733, (2006).
5. Lems, W. F. & Raterman, H. G.: Critical issues and current challenges in osteoporosis and fracture prevention. An overview of unmet needs. Ther Adv Musculoskelet Dis **9**, 299-316, (2017).
6. Clynes, M. A. et al.: The epidemiology of osteoporosis. British medical bulletin, (2020).
7. Fuggle, N. R. et al.: Fracture prediction, imaging and screening in osteoporosis. Nature Reviews Endocrinology **15**, 535-547, (2019).
8. Leslie, W. D. & Lix, L. M.: Comparison between various fracture risk assessment tools. Osteoporosis International **25**, 1-21, (2014).
9. Siris, E. S. et al.: Bone mineral density thresholds for pharmacological intervention to prevent fractures. Arch Intern Med **164**, 1108-1112, (2004).
10. Aspray, T. J.: Fragility fracture: recent developments in risk assessment. Ther Adv Musculoskelet Dis **7**, 17-25, (2015).
11. Papaioannou, A. et al.: 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. Cmaj **182**, 1864-1873, (2010).
12. Hippisley-Cox, J. & Coupland, C.: Predicting risk of osteoporotic fracture in men and women in England and Wales: prospective derivation and validation of QFractureScores. Bmj **339**, b4229, (2009).
13. Kanis, J. A., Johnell, O., Oden, A., Johansson, H. & McCloskey, E.: FRAX and the assessment of fracture probability in men and women from the UK. Osteoporos Int **19**, 385-397, (2008).
14. Nguyen, N. D., Frost, S. A., Center, J. R., Eisman, J. A. & Nguyen, T. V.: Development of prognostic nomograms for individualizing 5-year and 10-year fracture risks. Osteoporos Int **19**, 1431-1444, (2008).

연세대학교
YONSEI UNIVERSITY

15. Hollevoet, N., Verdonk, R., Kaufman, J. M. & Goemaere, S.: Osteoporotic fracture treatment. Acta Orthop Belg **77**, 441-447, (2011).

16. Force, U. P. S. T.: Screening for Osteoporosis to Prevent Fractures: US Preventive Services Task Force Recommendation Statement. JAMA **319**, 2521-2531, (2018).

17. Park, S. Y. et al.: Korean Guideline for the Prevention and Treatment of Glucocorticoid-induced Osteoporosis. J Bone Metab **25**, 195-211, (2018).

18. Kanis, J. A. et al.: FRAX and its applications to clinical practice. Bone **44**, 734-743, (2009).

19. Silverman, S. L. & Calderon, A. D.: The utility and limitations of FRAX: A US perspective. Curr Osteoporos Rep **8**, 192-197, (2010).

20. van Geel, T. A., Huntjens, K. M., van den Bergh, J. P., Dinant, G. J. & Geusens, P. P.: Timing of subsequent fractures after an initial fracture. Curr Osteoporos Rep **8**, 118-122, (2010).

21. Krupinski, E. A.: Current perspectives in medical image perception. Atten Percept Psychophys **72**, 1205-1217, (2010).

22. Madabhushi, A. & Lee, G.: Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis **33**, 170-175, (2016).

23. Frid-Adar, M. et al.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing **321**, 321-331, (2018).

24. Iqbal, T. & Ali, H.: Generative adversarial network for medical images (MI-GAN). Journal of medical systems **42**, 1-11, (2018).

25. Zhu, J., Yang, G. & Lio, P.: in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 1669-1673 (IEEE).

26. Creswell, A. et al.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**, 53-65, (2018).

27. Kazerouni, A. et al.: Diffusion models for medical image analysis: A comprehensive survey. arXiv preprint arXiv:2211.07804, (2022).

28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B.: in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10684-10695.

29. Ruiz, N. et al.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242, (2022).

30. Borges, J. L. C. et al.: Repeating Vertebral Fracture Assessment: 2019 ISCD Official Position. Journal of clinical densitometry : the official journal of the International Society for Clinical Densitometry **22**, 484-488, (2019).

연세대학교
YONSEI UNIVERSITY

31. Kong, S. H. et al.: Development of a Spine X-Ray-Based Fracture Prediction Model Using a Deep Learning Algorithm. Endocrinol Metab (Seoul) **37**, 674-683, (2022).
32. Hsieh, C.-I. et al.: Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning. Nature Communications **12**, 5472, (2021).
33. Dagan, N. et al.: Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. Nature Medicine **26**, 77-82, (2020).
34. Fleps, I. & Morgan, E. F.: A Review of CT-Based Fracture Risk Assessment with Finite Element Modeling and Machine Learning. Curr Osteoporos Rep **20**, 309-319, (2022).
35. Smets, J., Shevroja, E., Hügle, T., Leslie, W. D. & Hans, D.: Machine Learning Solutions for Osteoporosis—A Review. Journal of Bone and Mineral Research **36**, 833-851, (2021).
36. Hong, N. et al.: Deep-Learning-Based Detection of Vertebral Fracture and Osteoporosis Using Lateral Spine X-Ray Radiography. Journal of Bone and Mineral Research **38**, 887-895, (2023).
37. Wang, T. et al.: A review on medical imaging synthesis using deep learning and its clinical applications. Journal of applied clinical medical physics **22**, 11-36, (2021).
38. Goodfellow, I. et al.: Generative adversarial nets. Advances in neural information processing systems **27**, (2014).
39. Ho, J., Jain, A. & Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840-6851, (2020).
40. Tschuchnig, M. E. & Gadermayr, M.: in Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021. 33-38 (Springer).
41. Chen, L., You, Z., Zhang, N., Xi, J. & Le, X.: Utrad: Anomaly detection and localization with u-transformer. Neural Networks **147**, 53-62, (2022).
42. Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C.: in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. (eds Linwei Wang et al.) 35-45 (Springer Nature Switzerland).
43. Özbey, M. et al.: Unsupervised medical image translation with adversarial diffusion models. arXiv preprint arXiv:2207.08208, (2022).
44. Pinaya, W. H. et al.: in Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings. 117-126 (Springer).

연세대학교
YONSEI UNIVERSITY

45. Kodali, N., Abernethy, J., Hays, J. & Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215, (2017).

46. Kwon, M., Jeong, J. & Uh, Y. in International Conference on Learning Representations (2023).

47. Tumanyan, N., Geyer, M., Bagon, S. & Dekel, T.: Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. arXiv preprint arXiv:2211.12572, (2022).

48. Ferrar, L., Jiang, G., Schousboe, J. T., DeBold, C. R. & Eastell, R.: Algorithm-based qualitative and semiquantitative identification of prevalent vertebral fracture: agreement between different readers, imaging modalities, and diagnostic approaches. J Bone Miner Res **23**, 417-424, (2008).

49. Ferrar, L., Jiang, G., Clowes, J. A., Peel, N. F. & Eastell, R.: Comparison of densitometric and radiographic vertebral fracture assessment using the algorithm-based qualitative (ABQ) method in postmenopausal women at low and high risk of fracture. J Bone Miner Res **23**, 103-111, (2008).

50. Looker, A. C. et al.: Prevalence of low femoral bone density in older US adults from NHANES III. Journal of Bone and Mineral Research **12**, 1761-1768, (1997).

51. Tan, M. & Le, Q.: in International conference on machine learning. 6105-6114 (PMLR).

52. Li, H. et al.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47-59, (2022).

53. Dehouche, N. & Dehouche, K.: What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon, (2023).

54. Beaulieu-Jones, B. R. et al.: Evaluating Capabilities of Large Language Models: Performance of GPT4 on Surgical Knowledge Assessments. medRxiv, 2023.2007.2016.23292743, (2023).

55. Baumgartner, C.: The potential impact of ChatGPT in clinical and translational medicine. Clinical and translational medicine **13**, (2023).

56. Zhang, T., Irsan, I. C., Thung, F. & Lo, D.: Cupid: Leveraging ChatGPT for More Accurate Duplicate Bug Report Detection. arXiv preprint arXiv:2308.10022, (2023).

57. Sima, C. S. & Gönen, M.: Optimal Cutpoint Estimation With Censored Data. Journal of Statistical Theory and Practice **7**, 345-359, (2013).

58. Liu, X. & Jin, Z.: Optimal survival time-related cut-point with censored data. Stat Med **34**, 515-524, (2015).

59. Mills, E. S. et al.: Secondary Fracture Rate After Vertebral Osteoporotic Compression Fracture Is Decreased by Anti-Osteoporotic Medication but Not Increased by Cement Augmentation. J Bone Joint Surg Am **104**, 2178-2185, (2022).

연세대학교
YONSEI UNIVERSITY

60. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**, 837-845, (1988).

61. Wang, Z., Simoncelli, E. P. & Bovik, A. C.: in The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. 1398-1402 (Ieee).

62. Katzman, J. L. et al.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology **18**, 24, (2018).

63. Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A.: Evaluating the Yield of Medical Tests. JAMA **247**, 2543-2546, (1982).

64. Kamarudin, A. N., Cox, T. & Kolamunnage-Dona, R.: Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Medical Research Methodology **17**, 53, (2017).

65. Virtanen, P. et al.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261-272.

66. Esteban, C., Hyland, S. L. & Rätsch, G.: Real-valued (medical) time series generation with recurrent conditional gans. arXiv preprint arXiv:1706.02633, (2017).

67. Elazab, A. et al.: Macroscopic cerebral tumor growth modeling from medical images: A review. IEEE Access **6**, 30663-30679, (2018).

68. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. ArXiv **abs/2204.06125**, (2022).

69. Ray, P. P.: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems **3**, 121-154, (2023).

70. Liu, Y. et al.: Summary of ChatGPT-Related research and perspective towards the future of large language models. Meta-Radiology **1**, 100017, (2023).

71. Kim, S., Won, C. W., Kim, B. S., Choi, H. R. & Moon, M. Y.: The association between the low muscle mass and osteoporosis in elderly Korean people. J Korean Med Sci **29**, 995-1000, (2014).

72. Hong, S. & Choi, W. H.: The effects of sarcopenia and obesity on femur neck bone mineral density in elderly Korean men and women. Osteoporosis and Sarcopenia **2**, 103-109, (2016).

73. Go, S. W., Cha, Y. H., Lee, J. A. & Park, H. S.: Association between Sarcopenia, Bone Density, and Health-Related Quality of Life in Korean Men. Korean J Fam Med **34**, 281-288, (2013).

연세대학교
YONSEI UNIVERSITY

74. Al Saedi, A., Hassan, E. B. & Duque, G.: The diagnostic role of fat in osteosarcopenia. Journal of Laboratory and Precision Medicine **4**, (2019).

75. Lundberg, S. M. & Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30**, (2017).

76. Theis, L., Oord, A. v. d. & Bethge, M.: A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844, (2015).

77. Gui, J., Sun, Z., Wen, Y., Tao, D. & Ye, J.: A review on generative adversarial networks: Algorithms, theory, and applications. IEEE transactions on knowledge and data engineering **35**, 3313-3332, (2021).

78. Benny, Y., Galanti, T., Benaim, S. & Wolf, L.: Evaluation Metrics for Conditional Image Generation. International Journal of Computer Vision **129**, 1712-1731, (2021).

79. Shin, D. W., Cho, J., Park, J. H. & Cho, B.: National General Health Screening Program in Korea: history, current status, and future direction. Precis Future Med **6**, 9-31, (2022).

80. Stiglic, G. et al.: Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery **10**, e1379, (2020).

81. Paling, J.: Strategies to help patients understand risks. Bmj **327**, 745-748, (2003).

82. Swindell, J. S., McGuire, A. L. & Halpern, S. D.: Beneficent persuasion: techniques and ethical guidelines to improve patients' decisions. Ann Fam Med **8**, 260-264, (2010).

83. Geurts, E. M. A., Pittens, C. A. C. M., Boland, G., van Dulmen, S. & Noordman, J.: Persuasive communication in medical decision-making during consultations with patients with limited health literacy in hospital-based palliative care. Patient Education and Counseling **105**, 1130-1137, (2022).

84. Edwards, A., Elwyn, G. & Mulley, A.: Explaining risks: turning numerical data into meaningful pictures. Bmj **324**, 827-830, (2002).

85. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine **3**, 136, (2020).

86. Mitani, A. et al.: Detection of anaemia from retinal fundus images via deeplearning. Nature Biomedical Engineering **4**, 18-27, (2020).

87. Kanis, J. A. et al.: The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int **18**, 1033-1046, (2007).

88. Pisani, P. et al.: Major osteoporotic fragility fractures: Risk factor updates and societal impact. World J Orthop **7**, 171-181, (2016).

연세대학교
YONSEI UNIVERSITY

89. Kinoshita, H., Tamaki, T., Hashimoto, T. & Kasagi, F.: Factors influencing lumbar spine bone mineral density assessment by dual-energy X-ray absorptiometry: comparison with lumbar spinal radiogram. J Orthop Sci **3**, 3-9, (1998).

90. Zhang, B. et al.: Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. Bone **140**, 115561, (2020).

91. Sun, X. et al.: Prediction Models for Osteoporotic Fractures Risk: A Systematic Review and Critical Appraisal. Aging Dis **13**, 1215-1238, (2022).

연세대학교
YONSEI UNIVERSITY

Korean Abstract

언어 유도 척추 측면 X-ray 생성모델을 활용한 환자 맞춤형 골절위험도 예측 연구

<지도교수 홍남기, 김휘영>

연세대학교 대학원 융합의학과
조상욱

서론: 골다공증과 골절은 고령사회에서 사망률 및 이환률, 의료비용지출을 초래하는 주요질병 부담 중 하나이다. 효과적인 진단도구와 약물이 존재함에도 불구하고 치료를 받지 못하고 있는 환자가 40% 달하는 만큼 치료가 필요한 위험군을 효과적으로 스크리닝하여 낮은 진단율과 치료율에 대한 개선이 필요하다. 본 연구는 언어를 기반으로 가이드를 줄 수 있는 측면 X-ray 생성 모델을 구축하고, 생성된 이미지를 활용한 환자 맞춤형 골절 위험도를 예측하여, 골절 위험군을 나누어 개선 여부를 검정하였다.

연구방법: 모델구축 코호트는 2007년 1월 부터 2018년 12월까지 세브란스병원을 방문한 환자 중 측면 X-ray를 촬영한환자로 구성되었다. 미래에 발생된 척추 골절은 첫 방문 이후 추적관찰을 통해 정의되었다. 미래 이미지의 생성 모델은 두개의 확산모델로 구성되어 있고, 첫번째 모듈은 기존 이미지에서의 특징을 추출하고, 두번째 모듈에서는 언어 가이드를 입력 받아 이미지 생성한다. 초기 영상과 생성된 10년 후 영상에서 VERTE-X pVF 점수를 계산하고 (0~1점), 이를 기반으로 4개의

연세대학교
YONSEI UNIVERSITY

골절 위험군 (LL, LH, HIGH)으로 나누어 콕스비례위험모델을 활용한 골절 위험도를 예측한다.

결과: 총 9,276 명의 29,307장의 영상이 수집되었고(평균나이 65.7 세, 여성비율 66%), 평균 34.8 개월의 추적관찰기간동안 9.9%의 대상자에게서 골절이 발생되었다 (921 명 골절 발생). 추적관찰에서 얻은 영상과 만들어진 영상 간 pVF 값의 차이는 0.06 ± 0.2 로 0.655의 상관계수를 보이고 있다. LL그룹 대비 LH그룹, HIGH그룹에 속했을 때 골절발생위험도가 각각 109%, 391% 증가하였고 (위험비 [HR], 2.092, 4.911; P=0.001), 위험 그룹을 나이, 성별, 체질량지수로 이루어진 기본 모델에 추가하였을 때 모델의 적합도를 유의하게 개선 시켰다 (likelihood ratio 105.7, p <0.001). 골절 위험 그룹의 골절 예측력은 기존 FRAX 점수를 보정하였을 때에도 독립적으로 유의하였다. (위험비 [HR], 1.461; P=<0.001)

결론: X-ray와 기본 임상정보를 활용한 미래영상 생성과 pVF 점수는 개별화된 골절 위험도를 도출하고 추가비용없이 골절 예측력을 개선시킬 가능성이 있다.

핵심어: 골절위험도 예측, 골절, 확산모델, 딥러닝 예측 모델

연세대학교
YONSEI UNIVERSITY

PUBLICATION LIST

Hong, N., Cho, S.W., Shin, S., Lee, S., Jang, S.A., Roh, S., Lee, Y.H., Rhee, Y., Cummings, S.R., Kim, H. and Kim, K.M. (2023), Deep-Learning-Based Detection of Vertebral Fracture and Osteoporosis Using Lateral Spine X-Ray Radiography. J Bone Miner Res, 38: 887-895. https://doi.org/10.1002/jbmr.4814