diagnostics

MDPI

Article

Accuracy and Agreement of GPT-5.6 Sol-Generated Bone Density Estimates with Quantitative Computed Tomography Measurements: A Retrospective Study of 450 Adults

Omer Ozyigit [1,*], Elif Ozyigit [2], Serdar Serinsoz [1] and Hafize Uzun [3]

[1] Department of Radiology, Faculty of Medicine, Istanbul Atlas University, 34403 Istanbul, Turkey; serdar.serinsoz@atlas.edu.tr
[2] Department of Physical Medicine and Rehabilitation, Kanuni Sultan Suleyman Training and Research Hospital, University of Health Sciences, 34303 Istanbul, Turkey; dr.elifoz16@gmail.com
[3] Department of Medical Biochemistry, Faculty of Medicine, Istanbul Atlas University, 34403 Istanbul, Turkey; hafize.uzun@atlas.edu.tr
* Correspondence: omer.ozyigit@atlas.edu.tr

Abstract

Background and Objectives: Large language models (LLMs) are increasingly investigated for clinical decision support; however, their ability to predict quantitative densitometric parameters from non-imaging clinical information remains uncertain. This study assessed the accuracy, agreement, and exploratory discriminatory performance of GPT-5.6 Sol-generated lumbar and total hip T- or Z-score estimates against QCT-derived reference measurements. **Materials and Methods:** This retrospective, single-center study included 450 adults who underwent QCT examinations. Demographic, clinical, and laboratory data were provided to GPT-5.6 Sol using standardized patient-specific prompts without disclosing reference QCT measurements. Lumbar volumetric bone mineral density (vBMD) was assessed using the 3D Spine module, whereas proximal femoral measurements were obtained using the CTXA Hip module of QCT Pro^{TM} software. Each patient was evaluated once. Prediction performance was evaluated using Pearson's correlation coefficient (r), mean absolute error (MAE), root mean squared error (RMSE), Bland–Altman analysis, and receiver operating characteristic (ROC) curves. **Results:** The mean age was 62.02 ± 13.17 years, and 424 patients (94.2%) were women. GPT-5.6 Sol-generated scores showed modest correlations with measured values for the lumbar spine (r = 0.383) and total hip (r = 0.262). The corresponding MAE values were 1.256 and 0.941, and RMSE values were 1.550 and 1.266, respectively. Mean prediction bias was +0.339 for lumbar scores and −0.236 for total hip scores. For identifying CTXA-defined total hip osteoporosis (T-score ≤ −2.5), the area under the ROC curve (AUC) was 0.700 (95% confidence interval [CI]: 0.615–0.782), with 27.7% sensitivity and 90.2% specificity. Exploratory discrimination of lumbar vBMD < 80 mg/cm^3 yielded an AUC of 0.682 (95% CI: 0.626–0.738), with 61.2% sensitivity and 72.8% specificity. In an exploratory comparison, five-fold cross-validated linear regression yielded lower prediction errors than GPT-5.6 Sol across both anatomical sites and score-type subgroups. **Conclusions:** GPT-5.6 Sol demonstrated modest discrimination but limited individual-level accuracy in estimating QCT-derived bone density scores from clinical and laboratory information. Substantial prediction errors and low sensitivity for identifying CTXA-defined total hip osteoporosis indicate that GPT-5.6 Sol-generated estimates cannot substitute for direct densitometric measurements. Further investigations incorporating standardized prompting, reproducibility assessment, and external validation are warranted.

Check for updates

Academic Editor: Prabha Srinivasagam

Received: 25 October 2026
Revised: 12 November 2026
Accepted: 13 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: quantitative computed tomography; GPT-5.6 Sol; large language models; osteoporosis; bone mineral density; artificial intelligence; clinical prediction; CTXA

1. Introduction

Osteoporosis is a systemic skeletal disorder characterized by reduced bone strength, deterioration of bone microarchitecture, and increased susceptibility to fragility fractures. It represents a major public health concern, particularly among older adults and postmenopausal women, because osteoporotic fractures are associated with disability, reduced quality of life, and increased mortality. Dual-energy X-ray absorptiometry (DXA) remains the established clinical reference method for bone mineral density (BMD) assessment; however, limited screening and the absence of early symptoms may contribute to delayed diagnosis [1,2].

Quantitative computed tomography (QCT) provides complementary information on skeletal health by enabling three-dimensional assessment of volumetric BMD (vBMD), particularly in the trabecular compartment of the lumbar vertebrae. In addition, CT-based X-ray absorptiometry (CTXA) enables DXA-equivalent areal BMD assessment at the proximal femur. These approaches have expanded opportunities for osteoporosis evaluation, including opportunistic screening using existing CT examinations. Nevertheless, differences in imaging protocols, calibration methods, and anatomical reference standards remain important limitations [3,4].

Recent advances in artificial intelligence (AI) have facilitated automated bone density estimation, vertebral segmentation, and osteoporosis classification using CT images. Although imaging-based deep learning models have demonstrated promising performance, substantial methodological heterogeneity and limited external validation remain important challenges [5,6]. In parallel, machine learning approaches incorporating demographic, clinical, and biochemical variables have shown potential for identifying patients at increased risk of osteoporosis [2,7]. However, predicting categorical osteoporosis risk differs from accurately estimating continuous densitometric scores. Importantly, promising algorithmic performance does not necessarily establish clinical validity or real-world applicability. Acerra et al. [8], in a narrative review of artificial intelligence applications in implant dentistry, emphasized the need for rigorous clinical validation, external generalizability, and professional oversight before integrating AI-based systems into clinical decision-making.

Large language models (LLMs), including ChatGPT, have increasingly been investigated for clinical decision support. Previous research has evaluated their ability to provide osteoporosis-related recommendations consistent with clinical guidelines [9]. However, evidence regarding the ability of general-purpose LLMs to predict individual QCT-derived bone density scores directly from non-imaging clinical and laboratory information remains limited. Importantly, statistically significant associations between predicted and measured scores do not necessarily indicate sufficient agreement or accuracy for clinical application.

Therefore, the present retrospective study aimed to evaluate the predictive performance of GPT-5.6 Sol in estimating QCT-derived lumbar spine (L1–L4) and total hip T- or Z-scores using standardized demographic, clinical, and laboratory information from 450 adult patients. Reference measurements were obtained using the 3D Spine and CTXA Hip modules of QCT Pro^{TM} software, respectively, whereas GPT-5.6 Sol generated patient-specific predictions without access to the measured QCT results. The primary objective was to assess the accuracy and agreement of the predicted standardized scores with their corresponding QCT software-reported scores, rather than to predict measured volumetric or areal BMD directly. Secondary objectives were to investigate the ability of GPT-5.6 Sol-

generated scores to discriminate reduced bone density using anatomical site-appropriate reference criteria and to characterize potential limitations of LLM-based densitometric prediction. By evaluating numerical prediction performance rather than relying exclusively on categorical accuracy or general clinical responses, this study sought to provide evidence regarding the feasibility and limitations of using a general-purpose LLM for non-imaging-based estimation of skeletal density outcomes.

2. Materials and Methods

2.1. Study Design and Population

This retrospective, single-center observational study was conducted at Istanbul Atlas University Hospital, Istanbul, Türkiye. The study included 450 adult patients who underwent quantitative computed tomography (QCT) examinations between 2021 and 2025. Eligible participants were aged ≥ 18 years and had available demographic, clinical, laboratory, and QCT measurements. Patients younger than 18 years or with incomplete data required for the analyses were excluded.

Patient information was retrospectively retrieved from institutional medical records. Measured lumbar spine and total hip QCT-derived scores were used as reference outcomes to evaluate the predictive performance of GPT-5.6 Sol. The study was designed to assess the accuracy of predictions generated by a pretrained large language model rather than to develop or train a new machine learning algorithm.

2.2. Clinical and Laboratory Data

Demographic and clinical variables included age, sex, height, weight, body mass index (BMI), menopausal status, and previous fracture history. Laboratory parameters comprised serum 25-hydroxyvitamin D [25(OH)D], vitamin D status, calcium, phosphorus, parathyroid hormone (PTH), creatinine, and alkaline phosphatase (ALP).

These routinely available variables constituted the structured patient information provided to GPT-5.6 Sol. The model inputs did not include any radiological images, measured BMD values, or QCT-derived T- or Z-scores.

2.3. Quantitative Computed Tomography Assessment

Quantitative computed tomography (QCT) examinations were performed using a multidetector computed tomography scanner (Optima CT660; GE Healthcare Japan Corporation, Tokyo, Japan). All patients underwent QCT examinations using the same standardized institutional QCT 3D acquisition protocol. Scans were acquired with patients in the supine position, feet first, using a helical acquisition technique. The acquisition parameters included a tube voltage of 120 kVp, a tube current of 100 mA, a slice thickness of 2.5 mm, a reconstruction interval of 2.5 mm, and a gantry rotation time of 1.0 s. The gantry tilt was set to 0°, and the scan field of view was configured as Large Body.

QCT measurements were analyzed using QCT Pro^{TM} software (version 5.1.3; Mindways Software, Inc., Austin, TX, USA). Lumbar spine measurements were performed using the 3D Spine module, which provides three-dimensional trabecular volumetric bone mineral density (vBMD) measurements expressed in mg/cm^3. Trabecular regions of interest were assessed within the lumbar vertebral bodies. The lumbar vBMD values were analyzed as a distinct reference measurement and were not treated as numerically interchangeable with software-reported standardized scores.

Proximal femoral measurements were obtained using the CTXA Hip module, which generates dual-energy X-ray absorptiometry (DXA)-equivalent areal bone mineral density (aBMD) values expressed in g/cm^2, together with corresponding T- and Z-scores. Femoral neck and total hip bone mineral density values were also recorded.

A Mindways QA Phantom (HEAD; Part No. 13003; Mindways Software, Inc., Austin, TX, USA) was used for periodic calibration of the QCT system. The phantom was used for system calibration rather than being positioned beneath patients during individual examinations. Bone mineral density measurements were obtained directly from QCT Pro™ software-generated reports.

Lumbar spine and total hip T- or Z-scores were extracted directly from the software-generated reports without recalculation by the investigators. According to the study protocol, T-scores were used for postmenopausal women and men aged ≥ 50 years, whereas Z-scores were used for premenopausal women and men aged < 50 years. Measurements obtained from the 3D Spine and CTXA Hip modules were evaluated separately because of differences in measurement principles and reference standards. Accordingly, the term "QCT-derived score" in this study refers to a score reported by the relevant QCT analysis software, rather than to vBMD (mg/cm^3) or aBMD (g/cm^2) itself.

Representative images illustrating lumbar 3D Spine and proximal femoral CTXA Hip assessments are presented in Figure 1.

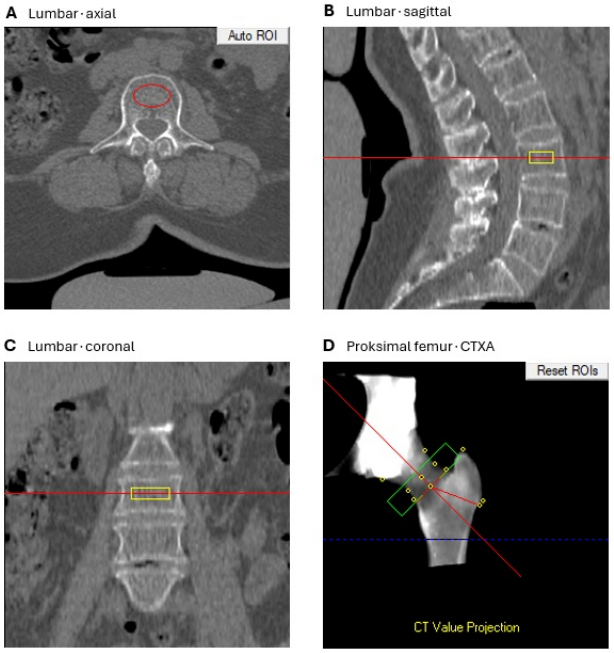


Figure 1. Representative quantitative computed tomography assessment of the lumbar spine and proximal femur. (**A**) Axial CT image depicting the trabecular region of interest within a lumbar vertebral body. (**B**) Sagittal and (**C**) coronal reconstructions illustrating vertebral localization for lumbar QCT assessment using the 3D Spine module. (**D**) Proximal femoral image illustrating the measurement regions used for CTXA Hip analysis.

2.4. GPT-5.6 Sol-Based Prediction Procedure

OpenAI's GPT-5.6 Sol large language model, accessed through the ChatGPT web interface (OpenAI, San Francisco, CA, USA), was used between 10 August and 30 September 2026. A standardized patient-specific prompting approach was applied consistently across all participants.

For each patient, the prompt included demographic, clinical, and biochemical information and instructed GPT-5.6 Sol to estimate lumbar spine (L1–L4) and total hip QCT-derived scores. The prompt explicitly specified that T-scores should be generated for postmenopausal women and men aged ≥ 50 years, and Z-scores for premenopausal women and men aged < 50 years.

The same standardized prompt structure was applied consistently to all 450 patients, with only patient-specific demographic, anthropometric, clinical, and biochemical values varying between assessments. Each patient was evaluated using a single query in a separate ChatGPT conversation. No preliminary prompt testing or iterative prompt refinement was performed before or during the evaluation of the study cohort. The reference QCT measurements were not disclosed to GPT-5.6 Sol during prediction generation. No systematically documented replicate predictions were available for estimating within-patient repeatability or response variability. The complete standardized prompt, translated from Turkish into English and presented with illustrative patient data, is provided in Supplementary Text S1.

No study-specific training, fine-tuning, or division into training, validation, and test datasets was performed. This design constitutes a fixed-model, single-prompt evaluation without model fitting on the study cohort; it does not provide external validation of a newly developed predictive algorithm.

2.5. Evaluation of Prediction Accuracy and Agreement

Continuous prediction performance was evaluated separately for lumbar spine and total hip scores. Associations between measured and predicted values were assessed using correlation coefficients. Prediction accuracy was quantified using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2).

Prediction error was defined as the predicted minus the measured score, with positive and negative values indicating overestimation and underestimation, respectively. Agreement was further evaluated using Bland–Altman analysis, including mean bias and 95% limits of agreement.

Because T-scores and Z-scores are standardized against different reference populations and have distinct clinical interpretations, prediction performance was assessed separately by score type. Pooled analyses combining T- and Z-scores were interpreted descriptively and not treated as a single clinically homogeneous endpoint.

2.6. Bone Density Classification and Discriminative Performance

Bone density classification analyses were conducted separately for lumbar spine and total hip measurements.

For lumbar QCT, measured trabecular vBMD was categorized as normal ($\geq$120 mg/cm^3), osteopenic range (80 to <120 mg/cm^3), or osteoporotic range (<80 mg/cm^3). The vBMD threshold of <80 mg/cm^3 defined the reference osteoporosis-range category, whereas a GPT-5.6 Sol-predicted T-score ≤ -2.5 was used as a prespecified exploratory classification cutoff. Because these measurements represent distinct scales, the thresholds were not considered numerically or diagnostically equivalent. Accordingly, the lumbar ROC analysis assessed discrimination of vBMD-defined low bone density rather than direct prediction accuracy for volumetric BMD.

For total hip CTXA measurements, conventional T-score thresholds were applied: normal (T-score ≥ -1.0), osteopenia ($-2.5 <$ T-score < -1.0), and osteoporosis (T-score ≤ -2.5).

In patients evaluated using Z-scores, values ≤ -2.0 were classified as below the expected range for age, whereas values > -2.0 were considered within the expected range for age.

Classification performance was assessed using confusion matrices, sensitivity, specificity, overall accuracy, and Cohen's kappa coefficient. Receiver operating characteristic (ROC) analyses were performed using continuous GPT-5.6 Sol-predicted scores, and discriminative performance was expressed as the area under the ROC curve (AUC).

For the Z-score subgroup, the reference outcome was defined as a measured lumbar spine or total hip Z-score ≤ -2.0. The minimum of the corresponding predicted Z-scores was used as the continuous ROC predictor, with its sign reversed so that higher values

indicated a greater likelihood of a positive reference outcome. Classification performance at the prespecified threshold was evaluated using a predicted minimum Z-score ≤ −2.0.

2.7. Statistical Analysis

Statistical analyses were performed using Python (version 3.13.5), with SciPy (version 1.17.0), NumPy (version 2.3.5), and scikit-learn (version 1.8.0). Continuous variables were summarized as mean ± standard deviation (SD), and categorical variables were expressed as frequencies and percentages.

The predictive performance of GPT-5.6 Sol was evaluated by comparing model-generated lumbar spine (L1–L4) and total hip T- or Z-scores with the corresponding QCT-derived reference measurements. Linear associations were assessed using Pearson's correlation coefficient (r). Prediction accuracy was quantified using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). R^2 was calculated as one minus the ratio of the residual sum of squares to the total sum of squares of the measured values. Negative R^2 values indicated poorer predictive performance than a constant prediction equal to the observed sample mean.

Differences between measured and GPT-5.6 Sol-predicted scores were assessed using paired-samples t-tests. Wilcoxon signed-rank tests were additionally performed as sensitivity analyses. Prediction error was defined as the predicted minus the measured score. Mean prediction bias and its 95% confidence interval (CI) were calculated from paired differences using the t-distribution. Agreement was further assessed using Bland–Altman analysis, with 95% limits of agreement calculated as the mean difference ±1.96 SD of paired differences. Continuous prediction performance was additionally evaluated separately in the T-score and Z-score subgroups.

Bone density classification and discriminative performance were assessed separately according to anatomical site and score type. Among patients evaluated using T-scores, total hip osteoporosis was defined as a measured CTXA T-score ≤ −2.5. For lumbar QCT, measured trabecular volumetric bone mineral density (vBMD) < 80 mg/cm^3 was used to identify the osteoporosis-range reference category. Discrimination of this outcome using GPT-5.6 Sol-predicted lumbar T-scores was considered exploratory because the predicted scores and reference vBMD measurements represent different scales. Among patients evaluated using Z-scores, values ≤ −2.0 were classified as below the expected range for age.

Receiver operating characteristic (ROC) curves were constructed using continuous GPT-5.6 Sol-predicted scores, and discriminative performance was quantified using the area under the ROC curve (AUC). The 95% CIs for AUC values were estimated using 1300 nonparametric bootstrap resamples, as specified in the original analysis. Threshold-based classification performance was evaluated using sensitivity, specificity, overall accuracy, confusion matrices, and Cohen's kappa (κ) coefficient. AUCs describe discrimination over varying score thresholds; reported sensitivity and specificity correspond to the prespecified classification thresholds rather than to an AUC-optimized threshold.

As an exploratory secondary analysis, conventional multivariable ordinary least-squares linear regression was compared with GPT-5.6 Sol using the same demographic, clinical, and laboratory variables (age, sex, height, weight, BMI, menopausal status, fracture history, 25-hydroxyvitamin D, vitamin D status, calcium, phosphorus, PTH, creatinine, and alkaline phosphatase). Separate models were fitted for lumbar spine and total hip scores within the T-score and Z-score subgroups. Regression performance was estimated using five-fold shuffled cross-validation (random seed 42), with preprocessing performed separately within each training fold, including median imputation and standardization for numerical variables and most-frequent imputation and one-hot encoding for categorical variables. Out-of-fold regression predictions were compared with the original GPT-5.6

Sol estimates on the same participants using MAE, RMSE, and R^2. This comparison was exploratory and did not constitute external validation.

All statistical tests were two-sided, and p-values < 0.05 were considered statistically significant.

2.8. Ethical Considerations

The study was approved by the Istanbul Atlas University Non-Interventional Scientific Research Ethics Committee (approval date: 26 January 2026; decision no. 01; meeting no. 01; official correspondence no. E-22686390-050.99-90293). The study was conducted in accordance with the principles of the Declaration of Helsinki. Patient information was retrospectively retrieved from institutional medical records and anonymized before analysis. No direct patient contact or intervention was involved. The requirement for informed consent was waived by the Ethics Committee due to the retrospective nature of the study and the use of anonymized patient data. Only anonymized, non-identifiable patient data were entered into the ChatGPT interface. No, personally identifiable information, including patient names, identification numbers, or other direct identifiers, was shared with the model.

3. Results

3.1. Study Population and Reference Bone Mineral Density Measurements

The analysis included 450 patients (mean age 62.02 ± 13.17 years; 424 women [94.2%]). T-scores were recorded for 372 patients and Z-scores for 78. Measured and GPT-5.6 Sol-predicted lumbar and total hip scores were available for all included patients. Mean lumbar trabecular volumetric bone mineral density (vBMD) was 97.57 ± 32.54 mg/cm^3; femoral neck and total hip areal BMD values were 0.71 ± 0.16 and 0.78 ± 0.14 g/cm^2, respectively (Table 1).

Table 1. Study characteristics and measured QCT-related outcomes.

Variable	Value
Number of patients	450
Age, years	62.02 ± 13.17
Female sex	424 (94.2%)
T-score/Z-score groups	372/78
Lumbar vBMD, mg/cm^3	97.57 ± 32.54
Femoral neck BMD, g/cm^2	0.71 ± 0.16
Total hip BMD, g/cm^2	0.78 ± 0.14
Lumbar vBMD < 80 mg/cm^3	135 (30.0%)
Lumbar vBMD 80 ≤ 120 mg/cm^3	218 (48.4%)
Lumbar vBMD ≥ 120 mg/cm^3	97 (21.6%)

Values are mean ± SD or n (%). vBMD denotes volumetric BMD; lumbar values were measured using the 3D Spine module, whereas hip aBMD values were derived from CTXA Hip.

Based on lumbar vBMD thresholds, 135 patients (30.0%) had values < 80 mg/cm^3, 218 (48.4%) had values from 80 to <120 mg/cm^3, and 97 (21.6%) had values ≥ 120 mg/cm^3. These categories were based on reference QCT measurements, not GPT-5.6 Sol-generated vBMD predictions.

3.2. Accuracy of Continuous Score Predictions

For the lumbar region, measured and predicted scores were −2.24 ± 1.56 and −1.90 ± 1.07, respectively. Pearson's r was 0.383, MAE 1.256, RMSE 1.550, and R^2 0.010. The

mean prediction error (predicted minus measured) was +0.339 (95% CI +0.199 to +0.479); the paired difference was statistically significant ($p < 0.001$).

For the total hip region, measured and predicted scores were −1.21 ± 1.15 and −1.44 ± 0.87, respectively. Pearson's r was 0.262, MAE 0.941, RMSE 1.266, and R^2 −0.223. The mean prediction error (predicted minus measured) was −0.236 (95% CI −0.352 to −0.121); the paired difference was statistically significant ($p < 0.001$) (Table 2).

Table 2. Measured versus GPT-5.6 Sol-predicted continuous QCT-derived scores (n = 450).

Site	Measured	Predicted	Pearson r	MAE	RMSE	R^2	Bias
Lumbar	−2.24 ± 1.56	−1.90 ± 1.07	0.383	1.256	1.550	0.010	+0.339
Total hip	−1.21 ± 1.15	−1.44 ± 0.87	0.262	0.941	1.266	−0.223	−0.236

Bias = predicted − measured. Scores comprise T- and Z-score observations; therefore, site-specific results do not represent one biologically homogeneous score scale. MAE, mean absolute error; RMSE, root mean squared error.

3.3. Agreement and Score-Type Stratification

Bland–Altman analysis showed a positive mean lumbar prediction bias and a negative mean total hip prediction bias. The 95% limits of agreement were wide at both sites (Table 3 and Figure 2), indicating substantial between-patient prediction discrepancies. Stratified continuous-score results are presented for the T-score and Z-score groups because these metrics use different reference populations.

Table 3. Prediction error and Bland–Altman limits of agreement overall and by score type.

Group	Region	n	MAE	RMSE	R^2	Bias	95% Limits of Agreement
Overall	**Lumbar**	450	1.256	1.550	0.010	+0.339	−2.63 to +3.31
Overall	**Hip**	450	0.941	1.266	−0.223	−0.236	−2.68 to +2.21
T-score	**Lumbar**	372	1.222	1.519	0.077	+0.157	−2.81 to +3.12
T-score	**Hip**	372	0.929	1.263	−0.233	−0.434	−2.76 to +1.89
Z-score	**Lumbar**	78	1.422	1.687	−1.155	+1.209	−1.11 to +3.53
Z-score	**Hip**	78	0.999	1.285	−0.377	+0.706	−1.41 to +2.82

95% limits of agreement = mean difference ± 1.96 SD of the differences. T-score and Z-score strata are reported separately to avoid conflating their distinct reference scales.

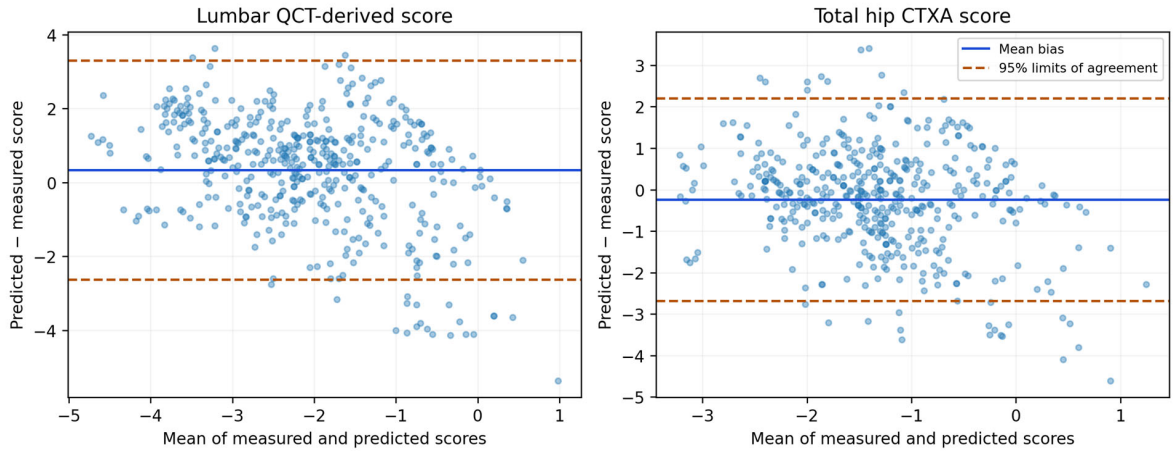


Figure 2. Bland–Altman plots comparing GPT-5.6 Sol-predicted and measured scores at the lumbar spine and total hip. Dashed lines indicate the 95% limits of agreement.

3.4. Anatomical Site-Specific Exploratory Discrimination

In the T-score subgroup (n = 372), 47 patients had total hip CTXA T-scores ≤ −2.5. Using the same threshold for GPT-5.6 Sol-predicted total hip T-scores, sensitivity was 27.7% and specificity 90.2%. Discrimination based on the continuous predicted total hip score yielded an AUC of 0.700 (bootstrap 95% CI 0.615–0.782) (Table 4).

Table 4. Exploratory discrimination and threshold-based classification by anatomical site and score type.

Outcome	Reference-Positive/n	ROC-AUC (95% CI)	Sensitivity	Specificity	Accuracy	Cohen κ
Hip CTXA T ≤ −2.5	47/372	0.700 (0.615–0.782)	27.7%	90.2%	82.3%	0.181
Lumbar vBMD < 80 (exploratory)	129/372	0.682 (0.626–0.738)	61.2%	72.8%	68.8%	0.331
Z-score ≤ −2.0	28/78	0.633 (0.503–0.764)	0.0%	100.0%	64.1%	0.000

ROC-AUC 95% CIs obtained from 1300 nonparametric bootstrap resamples (fixed random seed). Hip uses measured/predicted CTXA T-scores; lumbar uses measured vBMD and predicted lumbar T-score (different scales); Z uses the minimum of lumbar/hip measured or predicted Z-scores. Reference-positive denotes low score or low vBMD.

For the lumbar 3D QCT analysis, 129 of 372 T-score-group patients had measured trabecular vBMD < 80 mg/cm^3. GPT-5.6 Sol did not predict vBMD directly; therefore, discrimination using the predicted lumbar T-score was evaluated only as an exploratory cross-scale analysis. Its ROC-AUC was 0.682 (bootstrap 95% CI 0.626–0.738). At a predicted lumbar T-score threshold of ≤−2.5, sensitivity was 61.2% and specificity 72.8%; these threshold-based results should not be interpreted as validation of equivalent QCT and T-score diagnostic boundaries.

In the Z-score subgroup (n = 78), 28 patients had at least one measured lumbar or hip Z-score ≤ −2.0. The prediction-based classification yielded an AUC of 0.633 (bootstrap 95% CI 0.503–0.764). The corresponding sensitivity was 0.0% and specificity 100.0% at the prespecified score threshold. These results describe score-based agreement rather than QCT vBMD-defined osteoporosis. The ROC curves illustrating the discriminative performance of GPT-5.6 Sol-generated predictions for CTXA-defined total hip osteoporosis, low lumbar trabecular vBMD, and Z-scores below the expected range for age are presented in Figure 3. For the Z-score subgroup, ROC analysis was based on the negative of the minimum predicted lumbar spine and total hip Z-scores. The AUC reflects discrimination across all possible thresholds, whereas the reported sensitivity and specificity correspond to the prespecified Z-score cutoff of ≤−2.0.

All GPT-5.6 Sol-predicted minimum Z-scores were above −2.0 (range: −1.6 to +0.3). Therefore, none of the 28 reference-positive patients were identified, resulting in 28 false negatives, 50 true negatives, and 0% sensitivity. Despite this threshold-based failure, the continuous predictions provided modest discrimination (AUC = 0.633), reflecting differences in patient ranking across possible thresholds.

The corresponding confusion matrices, illustrating true-positive, true-negative, false-positive, and false-negative classifications for the three evaluated outcomes, are shown in Figure 4.

Cell values indicate the number of patients in each category. The lumbar analysis is exploratory because the reference classification was based on measured vBMD, whereas the predictions were based on GPT-5.6 Sol-generated T-scores.

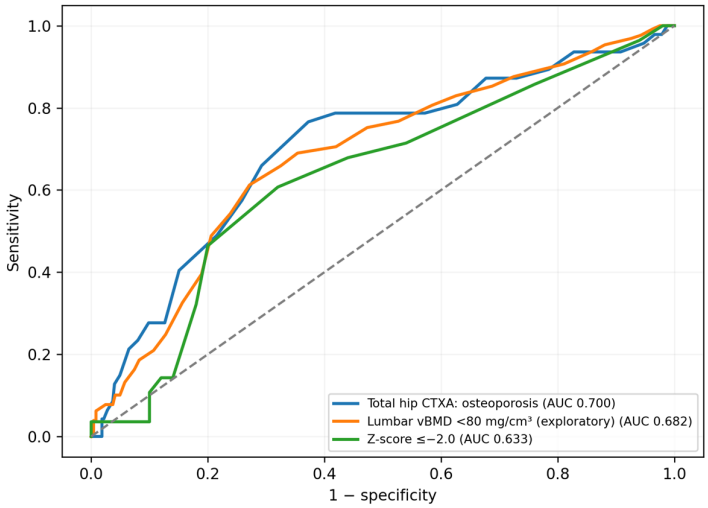


Figure 3. Receiver operating characteristic curves derived from continuous GPT-5.6 Sol predictions for total hip CTXA T-score osteoporosis, low lumbar trabecular vBMD, and low age-referenced Z-scores. Lumbar vBMD discrimination is exploratory across non-equivalent measurement scales. For the Z-score subgroup, the ROC-AUC was derived from the negative minimum predicted Z-score across the lumbar spine and total hip; threshold-based sensitivity and specificity were calculated at a predicted minimum Z-score of ≤−2.0.

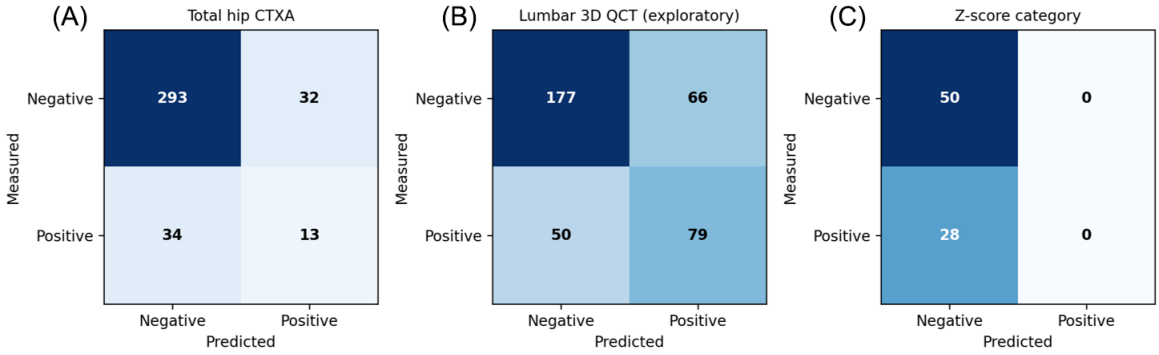


Figure 4. Confusion matrices comparing measured reference classifications with GPT-5.6 Sol-predicted classifications for (**A**) CTXA-defined total hip osteoporosis based on T-scores, (**B**) exploratory discrimination of lumbar trabecular vBMD below 80 mg/cm^3, and (**C**) Z-scores below the expected range for age. Rows represent reference classifications, whereas columns represent GPT-5.6 Sol-predicted classifications.

3.5. Comparison with Conventional Regression

In the T-score subgroup (n = 372), five-fold cross-validated linear regression yielded lower MAE than GPT-5.6 Sol for the lumbar spine (1.089 vs. 1.222) and total hip (0.815 vs. 0.929), with corresponding RMSE values of 1.451 vs. 1.519 and 1.102 vs. 1.263, respectively. In the Z-score subgroup (n = 78), regression also yielded lower MAE for lumbar spine (1.070 vs. 1.422) and total hip scores (0.918 vs. 0.999). Nevertheless, the regression R^2 remained negative for both Z-score outcomes (−0.390 and −0.225, respectively), indicating limited predictive adequacy. Detailed comparative metrics are presented in Table 5.

Table 5. GPT-5.6 Sol versus five-fold cross-validated linear regression.

Score Group	Site	Method	MAE	RMSE	R^2
T	Lumbar	GPT-5.6 Sol	1.222	1.519	0.077
T	Lumbar	Linear regression	1.089	1.451	0.158

Table 5. *Cont.*

Score Group	Site	Method	MAE	RMSE	R^2
T	Total hip	GPT-5.6 Sol	0.929	1.263	−0.233
T	Total hip	Linear regression	0.815	1.102	0.061
Z	Lumbar	GPT-5.6 Sol	1.422	1.687	−1.155
Z	Lumbar	Linear regression	1.070	1.355	−0.390
Z	Total hip	GPT-5.6 Sol	0.999	1.285	−0.377
Z	Total hip	Linear regression	0.918	1.212	−0.225

Values for regression represent pooled out-of-fold predictions from five-fold cross-validation; GPT-5.6 Sol estimates were previously recorded. Analyses were stratified by T-score and Z-score to avoid combining different reference populations.

4. Discussion

The principal finding of this study is that GPT-5.6 Sol demonstrated limited accuracy in predicting quantitative computed tomography (QCT)-derived bone density scores from routinely available clinical and laboratory information, despite statistically significant correlations with reference measurements. Among 450 patients, the correlations between measured and predicted scores were modest for the lumbar spine ($r = 0.383$) and total hip ($r = 0.262$), with mean absolute errors (MAEs) of 1.256 and 0.941, respectively. The low coefficient of determination for lumbar measurements ($R^2 = 0.010$) and negative value for total hip measurements ($R^2 = -0.223$) indicated poor individual-level predictive performance. For CTXA-defined total hip osteoporosis, GPT-5.6 Sol achieved an area under the receiver operating characteristic curve (AUC) of 0.700, with high specificity (90.2%) but low sensitivity (27.7%). Exploratory discrimination of lumbar volumetric bone mineral density (vBMD) below 80 mg/cm^3 yielded an AUC of 0.682, with sensitivity and specificity of 61.2% and 72.8%, respectively. Collectively, these findings suggest that although GPT-5.6 Sol may capture certain associations between clinical characteristics and skeletal density, its numerical predictions lack sufficient accuracy and sensitivity to substitute for direct densitometric assessment.

Ong et al. [10], in a systematic review of 39 studies, reported substantial variability in the diagnostic performance of CT-based AI models for osteoporosis classification (AUC: 0.582–0.994), highlighting methodological heterogeneity and limited external validation. Unlike these imaging-based approaches, the present study evaluated the ability of a general-purpose large language model to estimate QCT-derived bone density scores using only clinical and laboratory data. Although direct performance comparisons are inappropriate, both studies underscore the need for rigorous validation before clinical implementation. Recent advances in AI have demonstrated promising results in automated bone density assessment, particularly when imaging-derived information is incorporated into predictive algorithms. Paderno et al. [5], in a scoping review of AI-assisted opportunistic osteoporosis screening, highlighted the potential of CT-based AI applications while emphasizing methodological heterogeneity and the need for external validation. Similarly, Peng et al. [6] demonstrated the feasibility of deep learning-based bone density prediction and osteoporosis classification using CT images. Wu et al. [11] subsequently evaluated an automated deep learning system using CT examinations obtained from scanners produced by different manufacturers, demonstrating the feasibility of cross-scanner assessment. More recently, Sarquis Serpa et al. [12] reported multicenter validation of a deep learning algorithm for osteoporosis screening using abdominal CT images. These investigations support the potential clinical value of image-based AI systems. However, direct comparisons with the present study are inappropriate because GPT-5.6 Sol did not analyze patient-specific CT images, radiomic features, or vertebral attenuation measurements. Instead, its predictions were generated exclusively from structured clinical and laboratory information. Therefore,

our findings should be interpreted as an evaluation of non-imaging LLM-based prediction rather than an assessment of automated image interpretation.

The distinction between predicting osteoporosis risk and estimating continuous densitometric measurements is particularly important. Previous studies have demonstrated the potential utility of routinely available clinical and biochemical variables for identifying individuals at increased risk of osteoporosis. Liu et al. [2] developed a hierarchical screening approach incorporating clinical characteristics, laboratory information, and CT-derived imaging features. More recently, Wei et al. [7] developed an interpretable machine learning model using demographic characteristics and routine blood tests for osteoporosis prediction. Although these findings support the feasibility of data-driven risk stratification, predicting a categorical outcome differs fundamentally from estimating an individual patient's continuous T- or Z-score. The latter requires accurate approximation of the magnitude of skeletal density abnormalities rather than simply distinguishing patients at higher or lower risk. Our results suggest that the clinical and biochemical variables supplied to a general-purpose LLM, under the prompting conditions evaluated, were insufficient to reproduce individual QCT-derived scores with acceptable precision. Nevertheless, this finding should not be generalized to all non-imaging predictive models, particularly those specifically trained and validated for bone density estimation.

Anatomical site-specific differences represent another important aspect of the findings. GPT-5.6 Sol showed a higher correlation with measured lumbar spine scores than with total hip scores; however, substantial prediction errors were observed at both sites. Moreover, lumbar scores tended to be overestimated, whereas total hip scores were underestimated, suggesting systematic differences in the direction of prediction errors. These findings may partly reflect differences in anatomical composition, measurement techniques, and reference standardization. Lumbar assessments were performed using the 3D Spine module, which provides trabecular vBMD measurements, whereas proximal femoral assessments were performed using the CTXA Hip module, which generates DXA-equivalent areal bone mineral density measurements. Bott et al. [4] emphasized the effects of imaging protocols, calibration procedures, anatomical regions, and analytical methods on CT-based bone density measurements. Furthermore, Vermue et al. [13], in a systematic review involving 16 studies and 6772 participants, identified substantial variability in diagnostic thresholds and measurement procedures for opportunistic CT-based osteoporosis screening of the proximal femur. These observations support the separate evaluation of lumbar QCT and total hip CTXA outcomes and caution against assuming that their diagnostic thresholds are interchangeable.

The classification findings have clinically relevant implications, particularly regarding the identification of patients with osteoporosis-range measurements. Although GPT-5.6 Sol achieved 90.2% specificity for CTXA-defined total hip osteoporosis, sensitivity was only 27.7%, indicating that most patients meeting the reference criterion were not identified at the prespecified prediction threshold. Such a high false-negative proportion limits the potential usefulness of these estimates as a screening or exclusion tool. The lumbar analysis demonstrated somewhat greater sensitivity (61.2%), but overall discrimination remained modest (AUC = 0.682). Importantly, measured lumbar trabecular vBMD was used to define the reference outcome, whereas GPT-5.6 Sol generated standardized T-scores rather than direct vBMD estimates. Consequently, this analysis represents exploratory discrimination of low lumbar vBMD and should not be interpreted as direct validation of volumetric BMD prediction. These findings reinforce the importance of selecting anatomically and methodologically appropriate reference standards, as also emphasized by Mousavi et al. [3] and Vermue et al. [13]. In the Z-score subgroup, the absence of predicted values ≤ -2.0

resulted in complete failure to identify reference-positive cases at the prespecified threshold, despite modest discrimination based on continuous scores.

Another central observation is the discrepancy between statistical correlation and individual-level agreement. Although the associations between GPT-5.6 Sol-predicted and measured scores were statistically significant, their magnitudes were insufficient to establish clinical interchangeability. As emphasized by Mansournia et al. [14], correlation alone is insufficient to establish agreement between measurement methods, highlighting the importance of evaluating systematic bias and limits of agreement when assessing their potential interchangeability. The MAE and root mean squared error (RMSE) values indicated substantial individual prediction errors, while the low or negative R^2 values demonstrated limited predictive performance relative to the distribution of reference measurements. Bland–Altman analysis further characterized systematic bias and the dispersion of paired differences, providing a more informative assessment of agreement than correlation alone [15]. These observations are clinically relevant because errors near diagnostic thresholds may lead to incorrect classification. Accordingly, studies evaluating AI-generated numerical estimates should report absolute prediction errors and agreement measures rather than relying exclusively on correlation coefficients or classification accuracy.

The exploratory comparison with a conventional regression model provided additional context for interpreting GPT-5.6 Sol performance. Using the same clinical and laboratory predictors, cross-validated linear regression achieved lower absolute prediction errors at both anatomical sites and in both score-type subgroups. However, regression performance was also limited, particularly for Z-scores, suggesting that the available non-imaging variables may be insufficient for accurate individual-level densitometric estimation. These findings do not support a predictive advantage of the evaluated LLM prompting approach over a simple statistical comparator. Because the regression comparison was developed and internally cross-validated within this single-center cohort, independent external validation remains necessary.

The reliability of numerical predictions generated by large language models (LLMs) remains an important concern in clinical AI research. Gu et al. [16] investigated probabilistic predictions across multiple medical datasets and reported that explicitly generated probabilities may be less reliable than those derived from model output likelihoods. Although their study focused on probabilistic classification rather than continuous densitometric estimation, it highlights the need for rigorous evaluation of numerical outputs generated by LLMs. Similarly, Lin et al. [9] identified variability in the clinical appropriateness of recommendations provided by different LLMs for postmenopausal osteoporosis. Liu et al. [17] demonstrated the potential of ChatGPT-generated suggestions to optimize clinical decision support systems, while emphasizing limitations in their direct applicability. Furthermore, Dave et al. [18] highlighted broader concerns regarding the accuracy, reliability, transparency, and ethical implementation of ChatGPT in healthcare. Collectively, these studies underscore the importance of evaluating LLM-generated outputs against appropriate clinical reference standards. Extending this evidence to quantitative skeletal assessment, the present study demonstrated that GPT-5.6 Sol-generated estimates exhibited limited agreement with measured QCT-derived scores despite their numerical plausibility. These findings emphasize that the ability to generate patient-specific numerical predictions should not be equated with clinically acceptable predictive accuracy, and that rigorous validation is essential before such estimates can be considered for clinical decision support.

Reproducibility is another essential consideration. Franc et al. [19] investigated the accuracy, repeatability, and reproducibility of ChatGPT-based disaster triage and reported substantial variability across repeated assessments and different prompt formulations. Although that study examined a different clinical task and an earlier model, it demonstrates

the importance of systematically assessing output stability. In the present study, a standardized patient-specific prompt structure was used, and each patient was evaluated once. Informal repeat queries were performed for a small number of patients, but their outputs were not systematically documented. Consequently, within-patient prediction variability and reproducibility could not be quantified. Therefore, the present findings should be interpreted as reflecting the performance of GPT-5.6 Sol under the specific prompting conditions evaluated, rather than the general predictive reliability or reproducibility of LLM-based systems. Future studies should therefore incorporate repeated independent queries, prespecified prompt structures, and reproducibility analyses to determine the stability of LLM-generated densitometric estimates.

4.1. Strengths and Limitations

The present study has several strengths. It included 450 patients with paired measured and GPT-5.6 Sol-predicted scores, allowing quantitative assessment of prediction accuracy at two clinically relevant skeletal regions. The reference values were obtained directly from QCT Pro™ software-generated reports, and lumbar and proximal femoral measurements were evaluated using distinct analytical modules. Moreover, GPT-5.6 Sol predictions were generated without access to the measured QCT results, reducing the risk of direct reference-data leakage during prediction generation. The use of complementary statistical measures, including MAE, RMSE, R^2, Bland–Altman analysis, and ROC-AUC, enabled a more comprehensive evaluation than correlation or classification accuracy alone.

Nevertheless, several limitations warrant consideration. First, the retrospective, single-center design and predominantly female study population may limit generalizability. As women comprised 94.2% of the cohort, the findings may not adequately reflect model performance in male patients. The small number of men ($n = 26$) limited the feasibility of reliable sex-specific analyses, highlighting the need for validation in more sex-balanced cohorts.

Second, GPT-5.6 Sol was evaluated using a single standardized query per patient, precluding assessment of within-patient variability and test–retest reproducibility. Future studies should incorporate repeated independent queries to evaluate the consistency of LLM-generated densitometric estimates.

Third, the absence of study-specific model training or fine-tuning means that the results characterize the performance of the evaluated prompting approach rather than an optimized prediction model.

Fourth, although an exploratory internally cross-validated linear regression comparator was included, no independent external cohort, established clinical risk score comparison, or specifically trained machine learning comparator was available. Moreover, the relatively small Z-score subgroup may limit the stability of regression estimates. Consequently, these comparative findings should be regarded as preliminary and require external validation. Although conventional regression demonstrated lower prediction errors than GPT-5.6 Sol, the comparative analysis does not establish the superiority of either approach in independent populations. Further studies using external validation cohorts are warranted to assess their relative predictive performance and clinical utility.

Fifth, although the measured QCT results were not provided during prediction generation, the study could not independently establish whether comparable information had been encountered during the LLM's original pretraining.

Finally, the exploratory lumbar classification analysis compared predicted standardized scores with measured vBMD-defined categories, requiring cautious interpretation because these variables represent different measurement scales. Furthermore, the overall prediction error estimates combined T-scores and Z-scores derived from different reference populations; therefore, score-type-specific analyses provide a more appropriate basis for

clinical interpretation. Although comparative prediction errors were assessed, calibration and clinical utility were not formally evaluated.

Transparent reporting of these methodological features is particularly important for LLM-based clinical research. The TRIPOD-LLM reporting guideline proposed by Gallifant et al. [20] emphasizes detailed documentation of model characteristics, prompting procedures, outcome definitions, evaluation methods, and reproducibility. Although the present study standardized the patient information supplied to GPT-5.6 Sol, the lack of systematically recorded repeat-query results remains a limitation that should be addressed in future investigations.

4.2. Clinical Implications and Future Directions

The present findings do not support the use of GPT-5.6 Sol-generated QCT scores as replacements for directly measured densitometric outcomes. In particular, the low sensitivity for identifying CTXA-defined total hip osteoporosis raises concern regarding missed cases if such predictions were used for clinical screening or exclusion of reduced bone density. Nevertheless, the study contributes to the evaluation of general-purpose LLMs by examining their ability to generate numerical estimates from non-imaging information and identifying the limits of this approach. In particular, an apparently acceptable ROC-AUC cannot compensate for poor sensitivity at the clinically relevant prespecified decision threshold. Moderate discriminative performance should not be interpreted as sufficient clinical reliability, particularly when false-negative predictions may delay appropriate diagnostic evaluation and subsequent clinical management.

Future investigations should compare LLM-based predictions with conventional statistical models and dedicated machine learning algorithms using identical input variables and independent validation datasets. Additional studies should examine the effects of standardized prompting, repeated queries, model updates, and the inclusion of clinically relevant predictors. Multicenter prospective evaluations would also help establish whether prediction performance varies across patient populations and QCT acquisition systems. Until such evidence becomes available, LLM-generated densitometric estimates should be regarded as exploratory outputs rather than clinically validated measurements.

5. Conclusions

Overall, GPT-5.6 Sol produced patient-specific standardized bone density score estimates from routinely available clinical and laboratory information; however, individual-level agreement with software-reported reference scores was limited, and sensitivity for CTXA-defined total hip osteoporosis was low. The findings emphasize that generating plausible numerical predictions does not necessarily imply clinically meaningful quantitative accuracy, underscoring the need for rigorous methodological evaluation before LLM-based approaches are considered for densitometric decision support.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/diagnostics16233939/s1, Text S1: Standardized Patient-Specific Prompt Used for GPT-5.6 Sol Predictions.

Author Contributions: Conceptualization, O.O. and H.U.; methodology, O.O., E.O. and H.U.; software, O.O. and S.S.; validation, O.O., S.S. and H.U.; formal analysis, H.U.; investigation, O.O., E.O. and S.S.; resources, O.O. and S.S.; data curation, O.O. and S.S.; writing—original draft preparation, H.U. and O.O.; writing—review and editing, O.O., E.O., S.S. and H.U.; visualization, S.S. and H.U.; supervision, O.O. and H.U.; project administration, O.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Istanbul Atlas University Non-Interventional Scientific Research Ethics Committee (approval date: 26 January 2026; decision no. 01; meeting no. 01; official correspondence no. E-22686390-050.99-90293).

Informed Consent Statement: The requirement for informed consent was waived by the Istanbul Atlas University Non-Interventional Scientific Research Ethics Committee due to the retrospective nature of the study and the use of anonymized patient data.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data is not publicly available due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krueger, D.; Tanner, S.B.; Szalat, A.; Malabanan, A.; Prout, T.; Lau, A.; Rosen, H.N.; Shuhart, C. DXA Reporting Updates: 2023 Official Positions of the International Society for Clinical Densitometry. *J. Clin. Densitom.* **2024**, *27*, 101437. [CrossRef]
2. Liu, L.; Si, M.; Ma, H.; Cong, M.; Xu, Q.; Sun, Q.; Wu, W.; Wang, C.; Fagan, M.J.; Mur, L.A.J.; et al. A hierarchical opportunistic screening model for osteoporosis using machine learning applied to clinical data and CT images. *BMC Bioinform.* **2022**, *23*, 63. [CrossRef]
3. Mousavi, S.Z.; Moshfeghinia, R.; Vardanjani, H.M.; Sasani, M.R. Opportunistic screening of osteoporosis by CT scan compared to DXA: A systematic review and meta-analysis. *Clin. Imaging* **2025**, *118*, 110372. [CrossRef]
4. Bott, K.N.; Matheson, B.E.; Smith, A.C.J.; Tse, J.J.; Boyd, S.K.; Manske, S.L. Addressing Challenges of Opportunistic Computed Tomography Bone Mineral Density Analysis. *Diagnostics* **2023**, *13*, 2572. [CrossRef]
5. Paderno, A.; Ataide Gomes, E.J.; Gilberg, L.; Maerkisch, L.; Teodorescu, B.; Koç, A.M.; Meyer, M. Artificial intelligence-enhanced opportunistic screening of osteoporosis in CT scan: A scoping Review. *Osteoporos. Int.* **2024**, *35*, 1681–1692. Correction in *Osteoporos. Int.* **2024**, *35*, 1877. [CrossRef]
6. Peng, T.; Zeng, X.; Li, Y.; Li, M.; Pu, B.; Zhi, B.; Wang, Y.; Qu, H. A study on whether deep learning models based on CT images for bone density classification and prediction can be used for opportunistic osteoporosis screening. *Osteoporos. Int.* **2024**, *35*, 117–128. [CrossRef]
7. Wei, Q.; Zhan, J.; Chen, X.; Huang, Q.; Li, H.; Guo, W.; Liu, Z.; Chen, S.; Cai, D. Development and validation of an interpretable machine learning model for osteoporosis prediction using routine blood tests: A retrospective cohort study. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 446. [CrossRef]
8. Acerra, A.; Aliberti, A.; Amato, A.; Eccellente, A.; Santurro, A.; Giordano, F. Artificial Intelligence in Implant Dentistry: Clinical Validity, Diagnostic Performance, Surgical Planning, and Medico-Legal Implications-A Narrative Review. *Dent. J.* **2026**, *14*, 389. [CrossRef]
9. Lin, C.-R.; Chen, Y.-J.; Tsai, P.-A.; Hsieh, W.-Y.; Tsai, S.H.L.; Fu, T.-S.; Lai, P.-L.; Chen, J.-Y. Multiple large language models versus clinical guidelines for postmenopausal osteoporosis: A comparative study of ChatGPT-3.5, ChatGPT-4.0, ChatGPT-4o, Google Gemini, Google Gemini Advanced, and Microsoft Copilot. *Arch. Osteoporos.* **2025**, *20*, 120. [CrossRef]
10. Ong, W.; Liu, R.W.; Makmur, A.; Low, X.Z.; Sng, W.J.; Tan, J.H.; Kumar, N.; Hallinan, J.T.P.D. Artificial Intelligence Applications for Osteoporosis Classification Using Computed Tomography. *Bioengineering* **2023**, *10*, 1364. [CrossRef]
11. Wu, Y.; Yang, X.; Wang, M.; Lian, Y.; Hou, P.; Chai, X.; Dai, Q.; Qian, B.; Jiang, Y.; Gao, J. Artificial intelligence assisted automatic screening of opportunistic osteoporosis in computed tomography images from different scanners. *Eur. Radiol.* **2025**, *35*, 2287–2295. [CrossRef]
12. Sarquis Serpa, A.; Straus Takahashi, M.; Júdice de Mattos Farina, E.M.; Bennett, S.; Khandwala, N.; Major, M.; Paparian, J.; Garozzo Velloni, F.; Rothenberg, S.; Kayat Bittencourt, L.; et al. Advancing osteoporosis opportunistic screening: Multicenter validation of a deep learning algorithm using abdominal CT scans. *Abdom. Radiol.* **2026**, *51*, 2631–2641. [CrossRef]
13. Vermue, H.; Garibaldi, R.; Alshoaibi, Y.; Lustig, S.; Batailler, C. Opportunistic CT-based osteoporosis screening of the hip: A systematic review of diagnostic accuracy. *Arch. Orthop. Trauma Surg.* **2025**, *145*, 506. Correction in *Arch. Orthop. Trauma Surg.* **2025**, *145*, 524. [CrossRef]
14. Mansournia, M.A.; Waters, R.; Nazemipour, M.; Bland, M.; Altman, D.G. Bland-Altman methods for comparing methods of measurement and response to criticisms. *Glob. Epidemiol.* **2020**, *3*, 100045. [CrossRef]
15. Giavarina, D. Understanding Bland Altman analysis. *Biochem. Medica* **2015**, *25*, 141–151. [CrossRef]
16. Gu, B.; Desai, R.J.; Lin, K.J.; Yang, J. Probabilistic medical predictions of large language models. *npj Digit. Med.* **2024**, *7*, 367. [CrossRef]

17. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1237–1245. [CrossRef]
18. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef]
19. Franc, J.M.; Hertelendy, A.J.; Cheng, L.; Hata, R.; Verde, M. Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage of Simulated Patients Using the Simple Triage and Rapid Treatment (START) Protocol: Gage Repeatability and Reproducibility Study. *J. Med. Internet Res.* **2024**, *26*, e55648. [CrossRef]
20. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.