A COMPARISON OF INDICES FOR PREDICTING LOW BONE DENSITY IN POSTMENOPAUSAL CAUCASIAN WOMEN

by

Edward A. Mossman

A THESIS

Presented to the Department of Public Health and Preventive Medicine

and the Oregon Health & Science University School of Medicine

in partial fulfillment of the requirements for the degree of

Master of Public Health

April 30, 2002

School of Medicine
Oregon Health Sciences University

CERTIFICATE OF APPROVAL

This is certify that the MPH thesis of

Edward A. Mossman

has been approved

[redacted]
Thesis Committee Chair

[redacted]
Member

[redacted]
Member

Table of Contents

Acknowledgements

I am indebted to the members of my thesis committee for their guidance and assistance with this project:

Donald Freeborn, PhD

Jodi Lapidus, PhD

Michael McClung, MD

I would also like to acknowledge the assistance of the following people at the Oregon Osteoporosis Center for their outstanding commitment to the patients referred to our clinic, as well as to those of us who work with them there:

Nancy Grinnell

Lorna Cole

Kirsten Osborn

Terry DeFrancisco

Finally, I would like to extend my thanks to Merwyn "Mitch" Greenlick, PhD. This project benefited immensely from his and Dr. Freeborn's input during its inception and initial design.

Abstract

Osteoporosis leads to one and a half million fractures each year in this country. One out of every two women and one in eight men over 50 will have an osteoporosis-related fracture in their lifetime. Bone density testing by dual-energy X-ray analysis (DXA) is currently the best method available for diagnosing osteoporosis, but physicians are confronted with a wide and poorly defined array of clinical diagnostic features from which to choose in order to assess the value of referring patients for DXA testing. With the high cost of the DXA procedure to consider and a growing number of aging women potentially at risk, the use by physicians of an effective index scoring system based on readily available clinical factors could significantly reduce the number of patients unnecessarily referred for DXA testing. Several such indices are available at present, including the Simple Calculated Osteoporosis Risk Estimation (SCORE), Osteoporosis Risk Assessment Instrument (ORAI), and the Osteoporosis Self-Assessment Tool (OST). Quantitative ultrasound sonography (QUS) testing and similar methods of assessing bone at peripheral sites in the skeleton may represent a potential alternative screening methodology with which to select patients for DXA testing.

This study was performed to assess the three indices and compare them with one another on a large database of women referred for DXA testing at an osteoporosis clinic, to allow for a more complete picture of how useful such predictive schemes may be in a clinical setting. The primary objective of this study was to compare the statistical performance of the three different indices against each other and against quantitative ultrasound sonography in predicting low bone density in this population.

The four predictive schemes evaluated in this study proved similar in their ability to predict low bone density in this population, with QUS testing performing slightly better than the other predictors in diagnosing low bone density at the lumbar spine. Use of any one of these methods by primary care physicians could significantly reduce the number of women unnecessarily referred for bone density testing, at a potentially large cost savings to patients and health care organizations. Despite its slight predictive advantage at the spine, QUS testing is more expensive, time consuming, and requires an additional procedure compared to the predictive indices, which require only clinical information that is likely already obtained at the visit. The OST index requires only age and weight for calculation, and performs similarly to the other predictive schemes. The clinical indices may therefore be of greater clinical utility than QUS testing as a screening tool for referring patients for DXA.

A. Introduction/Specific Aims

Physicians are confronted with a wide and poorly defined array of clinical diagnostic features from which to choose in order to assess the value of referring patients for bone density testing by dual-energy x-ray analysis (DXA). Several indices have been created using different combinations of such clinically available data to evaluate patients' risk for low bone density, and therefore predict the necessity of DXA analysis. With the high cost of the DXA procedure to consider, and an ever-growing number of aging women potentially at risk, the use of an effective index by physicians could significantly reduce the number of patients unnecessarily referred for bone density testing, reducing the cost to patients and insurers, as well as potentially eliminating exposure to an additional radiographic test. Validation through population-based cross-sectional studies is ongoing, but it may be useful to evaluate how these indexes perform in a population referred for DXA testing to get a complete picture of how useful they may be in a clinical setting.

The primary objective of this study was to evaluate the performance (sensitivity, specificity, positive predictive value, negative predictive value, and area under the ROC curve) of three different indices for predicting low bone density in a large group of women referred by physicians for bone density testing. Indices were evaluated against each other and against quantitative ultrasound sonography (QUS) to determine comparative performance.

This study was designed to validate one or more of the three indices and compare them with one another on a large database of women referred for bone density testing at an

osteoporosis clinic. Comparisons to prior population-based development and validation studies of these indices, while not possible to evaluate statistically, are discussed. Additionally, it has been suggested that these indexes may perform equivalently or better than peripheral bone density testing in predicting hip and spine DXA. Part of the primary aim of this study was to evaluate the performance of QUS testing at the heel (Lunar Achilles, GE Lunar, Madison, Wisconsin) in this referral-based study population versus that of the predictive indices.

B. Background and Significance

Osteoporosis is characterized by a deterioration of bone density leading to an increased risk of fracture. These fractures can be severely debilitating, particularly in older individuals. According to the National Institutes of Health, more than 28 million Americans are affected by osteoporosis today, and the disease leads to a million and a half fractures each year in this country[1]. Almost one-third of patients with hip fractures are discharged to nursing homes within the year following a fracture, and one in five patients is no longer living 1 year after sustaining an osteoporotic hip fracture.

In a 1994 report, the World Health Organization provided a diagnostic definition of osteoporosis solely in terms of relative bone density values[2]. Specifically, osteoporosis was defined as a bone density value less than or equal to 2.5 standard deviations below the mean young adult normal value. This measurement system, using a reference population for comparison to young adult bone density values (referred to as "T-scores") as well as age-matched individuals (referred to as "Z-scores"), allows for evaluation of

both absolute and comparative fracture risk across a variety of populations and measurement methodologies.

Recent years have seen an explosion in the quantitative analysis of bone density by dual-energy x-ray absorptiometry (DXA), in which low-intensity X-ray photons are fired through tissue, collected, and then measured. Advances in the modality and frequency of bone density testing by DXA have been concomitant with similar growth in pharmacologic therapies for the treatment of low bone density, such as bisphosphonates, selective estrogen-receptor modulators (SERMS), and most recently parathyroid hormone. Despite these advances, most cases of osteoporosis in older women are still not diagnosed or treated by primary care physicians.[3] While the goal of bone density assessment by DXA is to identify patients at sufficient risk for future fractures to be candidates for such therapy, the time and money associated with screening every patient would be prohibitive.

With DXA costs at around $150 per scan, there has recently been a focus on determining the appropriate clinical factors to use in preliminary screening by physicians in primary care settings before referring patients for DXA testing.[4-10] Several indices have been derived for the purposes of determining appropriate candidates for DXA screening. These indices are based on criteria easily obtained in a clinical setting such as age, weight, and fracture history, and would allow a physician to determine which patients would be good candidates for DXA screening while eliminating from consideration those patients for whom it is likely unnecessary. Table 1 summarizes the indices examined in this study, and they are discussed in detail below.

The SCORE index

The first of these indices to be developed was the Simple Calculated Osteoporosis Risk Estimation (SCORE), which was first published in 1998 by Lydick et al[11], and has subsequently had mixed results in validation studies.[12-19] It uses a weighted combination of age, any previous or current use of estrogen replacement therapy, fracture history, race, weight, and history of rheumatoid arthritis to determine a score. If the number arrived at is 6 or greater, then the patient would be considered a candidate for DXA testing by this index.

The ORAI index

The second index that will be evaluated in this study, the Osteoporosis Risk Assessment Instrument (ORAI), was published in 2000 by Cadarette et al[20], and uses age category, weight, and current estrogen use to calculate a diagnostic score. Patients scoring 9 or greater on this index would be recommended for DXA referral.

The OST index

The most recently developed of the indices is the Osteoporosis Self-assessment Tool (OST). It was published in 2001 by Koh et al[21], and is the simplest of all three indices to use. It was developed on a population of Asian women, and it will be of particular interest to determine if it performs as well as the other two indices in this study, which were developed on largely Caucasian study populations. The OST uses only age and

weight to calculate a score. Patients who scored –1 or lower would be candidates for DXA screening according to this test.

Quantitative Ultrasound Sonography (QUS)

Several technologies have emerged as alternatives to DXA for bone density testing. Many of them measure sites on the peripheral skeleton, such as finger, wrist or forearm, and heel using X-rays or ultrasound attenuation to measure bone density. The value of peripheral bone density in predicting fracture risk, and its relationship to DXA measurements of the axial skeleton, remains a topic of debate. [22-27] One use of peripheral bone density measuring devices gaining popularity is as a screening tool to assist physicians in primary care settings to determine candidates for further testing. The most commonly used methods of peripheral bone density measurement include DXA at the forearm or finger, and QUS or Single Photon X-Ray Absorptiometry (SXA) at the heel. Of these different methods, heel ultrasound (QUS) testing appears to best predict fracture risk at common T-score thresholds.[28] This study examined heel ultrasound measurements by the Lunar Achilles (GE Lunar, Madison, Wisconsin) in the study population to determine if the simple clinical indices performed as well as the QUS procedure in predicting low bone density by DXA. While a T-score of –2 or –2.5 is commonly used as a cutoff in DXA testing for classification of low bone density, there is no consensus as to the best cutoff to demarcate those patients with elevated fracture risk. In this study, the QUS T-score for comparison was fixed in a post-hoc fashion by determining the threshold value that best matched the sensitivity levels of QUS with those of the clinical indices in predicting DXA T-scores of –2 and –2.5.

Table 1. Summary of four predictive schemes for low bone density.

Index/Test Name	Criteria Used	Threshold for LBD prediction
Simple Calculated Osteoporosis Risk Estimation (SCORE)	Age: 3 x first digit of age (e.g., 57 yo = +15) Estrogen Replacement therapy (ever): +1 if No Low-trauma fractures of wrist, hip, or rib after age 45 (+4 each to +12 max) Race: +5 if not black Rheumatoid arthritis: +4 if yes Weight: Subtract (wt in lbs ÷ 10), rounded down to nearest whole number	≥ 6
Osteoporosis Risk Assessment Instrument (ORAI)	Age: ≥75 yo + 15; 65-74 + 9; 55-64 +5 Estrogen Replacement therapy (current): +2 if No Weight: 60-69.9 kg +3; <60 kg +9	≥ 9
Osteoporosis Self-Assessment Tool (OST)	0.2 x (Weight-age), rounded down to nearest whole number	≤ -1
Qualitative Ultrasound Sonography (QUS)	T-Score (standard deviations from young normal value)	≤ -1

Other predictive schemes

Excluding QUS, the indices above were chosen for their ease of use and relatively small number of predictive factors to identify. Other predictive mechanisms and guidelines have been suggested for low bone density and/or DXA testing, however were inappropriate for inclusion in this study for various reasons. First, there exist guidelines published by organizations such as the National Osteoporosis Foundation, the International Society for Clinical Densitometry, and a myriad of health systems interested in limiting their providers' referrals for expensive procedures such as DXA. Such guidelines were not included in this project because, by and large, they still are intended to rely in varying degrees on the subjective judgment of the physician to evaluate the appropriate level of future fracture risk in each case, and therefore the necessity of DXA

screening. Other fracture-based predictors attempt to create a fracture risk profile that includes factors unrelated to bone density, and therefore are not evaluable using DXA testing alone as the endpoint. With regard to other predictive indices that do not rely on subjective assessment, the investigator is aware of only two other such published indices. One, known as SOFSURF, was not included in this study because it has been published as an abstract only since its design in 1998. It apparently did not at that time appear to show sufficient predictive ability to justify its further publication, and for similar reasons it is not included in this project. The other, known as ABONE, uses essentially the same variables as the ORAI but with a different scoring system marked by a lower degree of stratification by age group, and -- possibly as a result -- has tested poorly in validation studies. Given the existence of the ORAI index with the same variables and much better predictive ability, the ABONE index was left out of this study as well.

The objective of this study is to evaluate the three indices and QUS testing on a large database of women referred to an osteoporosis clinic by primary care physicians, in order to evaluate their sensitivity and specificity in predicting low bone density. DXA results of 2 and 2.5 standard deviations below the young normal mean (T-scores of –2 and –2.5) will be used to define low bone density at both the lumbar spine and femoral neck regions. Positive predictive value, negative predictive value, and area under the ROC curve will also be examined to evaluate each of the indices.

Due to the referral-based nature of this study population, comparisons with the results of validation studies on each of these indices should provide some insight into the ability of physicians to select women for bone density testing by DXA. Since it is likely easier for

physicians to be inclusive rather than exclusive in using subjective, off-hand criteria or loose guidelines to determine candidates for bone density screening, this study will pay particular attention to the comparison of specificity among the referral-selected women comprising this group with that of the population-based studies. Sensitivity represents the probability that a test would predict normal DXA bone density (and therefore no need for referral) given that the patient's DXA T-score was indeed above the endpoint threshold, and therefore indicates how many of those who did not need DXA screening would be 'weeded out.'

C. Study Design and Methods

1. Overview

This study was designed to compare the predictive ability of several screening tests in a cross-sectional fashion. From February 2000 through March 2002, approximately 800 referrals for bone density testing at the Oregon Osteoporosis Center, LLC, were part of an ongoing project to evaluate the clinical utility of heel ultrasound as a part of the Center's diagnostic options. Postmenopausal Caucasian women referred for bone density testing at the Center, without common secondary causes of osteoporosis (such as steroid use), with no history of antiresorptive therapy other than hormone replacement, and with no signs of degenerative joint disease (DJD) on the DXA scan were selected by technicians as time permitted to have heel ultrasounds (QUS) performed using a Lunar Achilles sonimeter (GE Lunar Corporation, Madison, Wisconsin) in addition to the typical hip and spine DXA measurements. All patients receiving QUS testing during this effort had their

DXA measurements performed on one DXA system, a Lunar Prodigy (GE Lunar, Madison, Wisconsin).

Clinical information required to calculate scores for each of the predictive indices was obtained from the Center's electronic medical record system. This system stores data in a Microsoft SQL Server 2000 database and contains responses to the standard patient questionnaire as well as patient scan results for both DXA and QUS measurements. Scores for each of the indices were calculated from the questionnaire data and compared with each other and with QUS results to examine their ability to predict DXA scores of 2 and 2.5 standard deviations below young normal values.

2. Subjects

a. Data Sources

Data gathered from patients in the study population were extracted from the Oregon Osteoporosis Center electronic medical records database, which contains clinical data from patient health status/history questionnaires mailed to all patients referred to the Center for bone density testing, as well as scan results from DXA and QUS testing for the entire population. The data in this system is maintained in a Microsoft SQL-Server 2000 database, with local tables on client systems in Microsoft Access format.

b. Inclusion Criteria

Postmenopausal Caucasian women with no known secondary causes of osteoporosis, who had QUS measurements performed as part of their bone density referral procedures at the Oregon Osteoporosis Center from 2/15/2000 to 3/26/2002 were included in the

analyses. QUS testing relies on comparative scoring against a reference population rather than absolute values, and while the Lunar Achilles sonimeter is among the most studied peripheral bone density instruments, its reference population is comprised and validated only among postmenopausal Caucasian women. For this reason, technicians were instructed to perform QUS scans only on Caucasian women. The attending technicians were also instructed to exclude from QUS testing women with a self-reported history of antiresorptive therapies other than hormone replacement therapy, as well as women with radiographic or self-reported evidence of degenerative disease or scoliosis, which could lead to overestimation of DXA spine bone density values. Valid entry of the patient's pre-scan questionnaire for the visit into the electronic medical records system was also required to be included in the analysis. This questionnaire, distributed to all referral patients to be completed before their scans, contains information about exclusion criteria for this study, such as antiresorptive therapies or steroid use, as well as data required to calculate components of the SCORE and ORAI indices, such as HRT use, fracture history, and rheumatoid arthritis.

3. Data Collection

a. Exclusions/Patient Accounting

From the electronic medical records, 814 women were identified as having both a record of a QUS scan and a valid questionnaire entry at a referral visit between 2/15/2000 and 3/26/2002. Three records were excluded for having no values entered in the ethnicity field, 29 indicated that a diagnosis of osteoporosis had been made by DXA prior to the referral visit, and 57 were not identified as postmenopausal. Twenty-one records were

excluded due to current steroid use. Some overlap existed among these groups. Out of the 709 remaining records, 39 were excluded for use of antiresorptive therapies including Didronel (1), raloxifene/Evista (9), alendronate/Fosamax (24), and calcitonin (5). Three records were excluded for current use of seizure medications, and one for an eating disorder. Six hundred sixty-six patients remained at this stage, and 5 more were subsequently removed for unknown estrogen replacement status, leaving 661 patients for the analysis. Five of the remaining 661 records included in the analysis did not have femoral neck bone density values at the visit (possibly due to bilateral hip replacement or degenerative disease), and 2 others did not have spine DXA values.

Table 2: Summary of records excluded from analysis

Exclusion Reason	**Number**	**Exclusion Reason**	**Number**
No information on ethnicity*	3	Antiresorptive therapy†	39
Prior diagnosis of OP by DXA*	29	Current seizure medications	3
Not specified as postmenopausal*	57	Eating disorder	1
Current steroid use*	21	No information on ERT status	5

Initial number of records: 814 Final number of records analyzed: 661
5 of remaining 661 records included only in spine endpoint analyses, 2 included only in hip endpoints.
2 patients left out of SCORE analyses due to incomplete information

*Some patients overlap between groups;
†Didronel (1), Calcitonin (5), Raloxifene (9), Alendronate (24)

b. Data Management/Cleaning

Clinical information on age, weight, and other data required to calculate the values of clinical indices were assembled in a Microsoft Access database, where the patient name, demographic information not required to calculate the indices, and any other personally identifiable information was removed for confidentiality. Scores for each of the clinical indices were then calculated for each patient and stored as separate variables. New fields were also created to denote the prediction by each index of low bone density, as Yes/No,

using the threshold scores determined by the authors. Data were then exported to other software as necessary for portions of the statistical analysis.

Information for the calculation of the predictive indices was compiled and cleaned according to the component requirements of each index. The SCORE index includes an adjustment for each low-trauma fracture of the wrist, hip, or rib after the age of 45, and information on fracture history was compiled from questionnaire data in order to calculate this adjustment. All patient fractures occurring before the age of 45, or where the age was reported as unknown, were removed from these calculations, as were those noted as resultant to a motor vehicle accident or other high levels of trauma. Fractures with a physical locale reported as "hand" were included with wrist fractures in the scoring, while those reported as "finger" were not. Fracture locales reported as unknown were also not included in calculating SCORE index results.

Another component of the SCORE index requires knowledge of any previous estrogen replacement therapy. Two patients did not report current ERT use, but had unknown previous ERT status. These patients were left out of the analyses for all comparisons involving the SCORE index.

4. Measurement of Variables

a. Outcome Variables

DXA bone density at the femoral neck (FN) and lumbar spine (LS) constituted the outcome variable, measured in standard deviations from the young normal value (T-score). The outcome threshold values considered for comparison of the predictive value

of the indices and QUS testing were 2 and 2.5 standard deviations below young normal. As discussed in the background and significance section, The World Health Organization has defined osteoporosis as a T-score of less than -2.5, but there is some debate as to how this translates to fracture risk, which is a major reason for assessing bone density. The North American Menopause Society recommends treatment for postmenopausal women under the age of 65 and with no other risk factors for osteoporosis if they have a femoral neck DXA T-score less than or equal to -2,[29] and for all postmenopausal women with hip or spine T-scores less than or equal to -2 and at least one additional risk factor for fracture.[30] T-score cutoffs of both -2 and -2.5 will therefore be used as endpoints at both femoral neck and lumbar spine sites, and because clinicians consider women who may have low bone density at either site for DXA testing, the additional outcome of either femoral neck or lumbar spine bone density below the respective thresholds of -2 and -2.5 was also examined in the analysis.

b. Predictor Variables

Predictor variables were created by calculating scores for each record according to the three indices, and determining if each met the threshold value of the index for prediction of low bone density. The summary of scoring systems for each index is detailed in the background section.

Three dichotomous variables, indicating whether each index predicted low bone density by DXA, were created in addition to the continuous variables containing the resultant score from each index. QUS testing was included as a fourth predictive scheme, both as a continuous variable reflecting the patients' QUS T-Scores, and as a dichotomous

variable counting threshold T-scores less than or equal to –1 as predictive of low bone density. The rationale for this cutoff will be described in detail in the results.

5. Statistical Analysis

Spearman correlations were performed on each possible pair of predictors using absolute index scores and QUS T-scores, independent of any threshold cutoffs. Scatterplots were made for each comparison. Spearman's method is more robust to nonparametric data than Pearson's method of correlation, which assumes normality. This method was used due to the potential of the different weights assigned to components in the index scoring schemes to produce a skewed distribution of scores.

Several different statistical methods were used to compare the predictive and selective abilities of the clinical indices and QUS testing across a total of 6 separate outcomes, representing DXA bone density T-scores less than or equal to -2 and -2.5 at the femoral neck, lumbar spine, and at either femoral neck or lumbar spine. For each endpoint, calculations were made of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the ROC curve. The table and figure below summarize these measures and how they were calculated.

Figure 1. 2x2 table of possible test/disease outcomes

	Disease Positive	**Disease Negative**	
Test Positive	True Positives (A)	False Positives (B)	*A+B: Total # of positive tests*
Test Negative	False Negatives (C)	True Negatives (D)	*C+D: Total # of negative tests*
	A+C: Total # with Disease	*B+D: Total # without Disease*	*A+B+C+D: Total # of subjects*

Table 3: Summary of statistical measures used.

Statistical Comparison	Definition	Dependent on Test Threshold?	Formula (see figure above)
Sensitivity	Probability of testing positive given disease present	Yes	A/(A+C)
Specificity	Probability of testing negative given disease absent	Yes	D/(B+D)
Positive Predictive Value (PPV)	Probability of disease presence given positive test	Yes	A/(A+B)
Negative Predictive Value (NPV)	Probability of disease absence given negative test	Yes	D/(C+D)
Area under ROC Curve (AUROC)	Total area under plotted curve of sensitivity vs. (1-specificity)	No	n/a

Sensitivity, specificity, PPV, and NPV all rely on a threshold value to be used as a diagnostic threshold for predictor variables. These thresholds split the continuous index score (or T-score value, in the case of QUS) into a dichotomous variable predicting either presence or absence of the outcome. Each predictive index uses a threshold cutoff determined by its creators to predict low bone density, and a cutoff T-score for QUS testing of less than or equal to -1 was established in a post hoc fashion by setting the sensitivity roughly equivalent to those of the other predictors. This obviously makes comparison of QUS sensitivity to the other predictors of little value, but does allow for a comparison of the other properties. The area under the ROC curve (AUROC) represents the probability that two randomly selected cases, one each from the abnormal/diseased population and the normal/non-disease population, will place in the correct order on the predictive scale being examined (e.g., the abnormal case will score higher than the lower one).[31] Measurement of AUROC has the advantage of being threshold independent, in that it calculates a probability by taking a number of thresholds and plotting the resultant sensitivity and one minus the resultant specificity. For this reason, measurement and comparison of AUROC is of central importance in evaluating the performance of the indices.

Comparative tests devised by Hanley and McNeil[32] (Analyse-it Clinical Laboratory 1.62, Analyse-it, Ltd., Leeds, England, UK) and Delong, Delong, and Clarke-Pearson[33] (AccuROC 2.5 for Windows, Accumetric Corp., Montreal, PQ, Canada) were used to test the statistical significance of the difference between AUROC scores among the indices and QUS at each endpoint. Generalized estimating equations (GEE) were used in a binary logistic model[34, 35] (SUDAAN 8.0, Research Triangle Institute, Research Triangle Park, NC) to test the significance of the differences between the respective sensitivity, specificity, PPV and NPV of each predictor at each outcome. GEEs provide a method of comparing binary outcomes in studies involving correlated data, such as repeated measures designs. In this study, GEEs compared conditional probabilities from different diagnostic tests performed on the same cases. For example, to test differences in sensitivity, GEEs were used to compare the proportion of patients predicted to have low bone density between the different predictors, after selecting only cases whose bone density by DXA was indeed at or below the given endpoint threshold. Likewise, differences in PPV were tested by comparing the proportion of subjects with low DXA bone density after selecting on cases who were predicted to have low bone density by the respective predictors. In these GEE comparisons, each predictive scheme is represented as a different level of a single variable, so comparison of each of the four against each other (6 comparisons) was made using a p-value of 0.00833 after applying the Bonferroni adjustment for multiple comparisons. (0.05/6 = 0.0083).

D. Results

Correlations

In comparing the indices with each other and QUS testing, it is of some use to examine the correlations between each predictor individually. Below in figures 2-7 are six graphs plotting each combination of predictors for all cases used in this study. In each graph, the predictive thresholds of the given variables are marked with perpendicular lines, and an 'X' marks the quadrant in which both variables would predict low DXA bone density. Empirically, the graphs indicate that the indices correlate better with each other than with QUS testing. The OST and SCORE indices appear to have the tightest representation of these plots, followed by the ORAI vs. OST and ORAI vs. SCORE comparisons. As table 4 shows, Spearman correlations show that this is indeed the case.

Figure 2. Scatter Plot of ORAI vs. SCORE

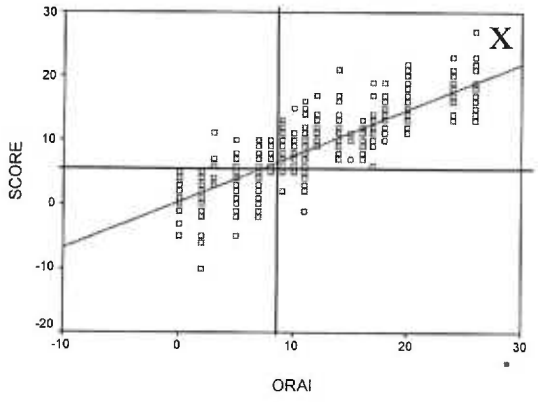

Figure 3. Scatter Plot of QUS vs. SCORE

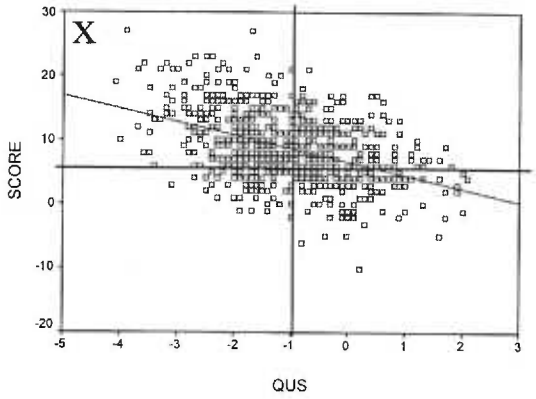

Figure 4. Scatter Plot of OST vs. SCORE

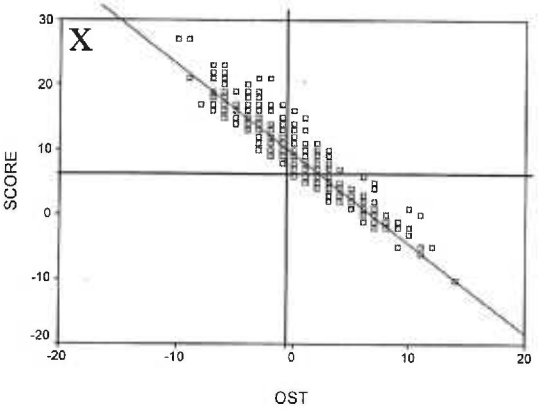

Figure 5. Scatter Plot of ORAI vs. OST

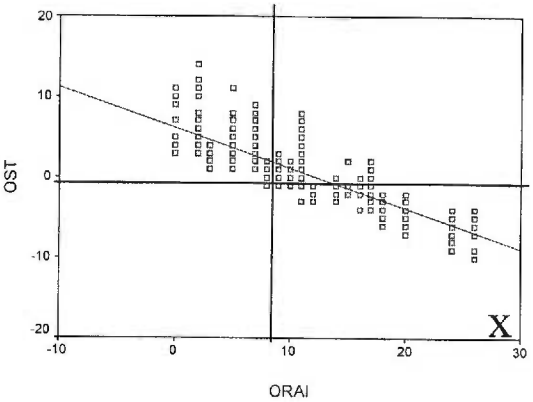

Figure 6. Scatter Plot of ORAI vs. QUS

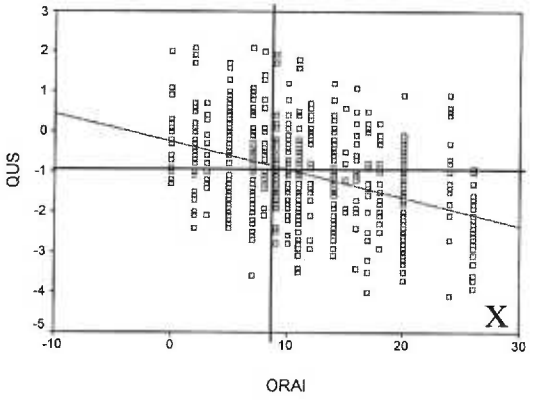

Figure 7. Scatter Plot of QUS vs. OST

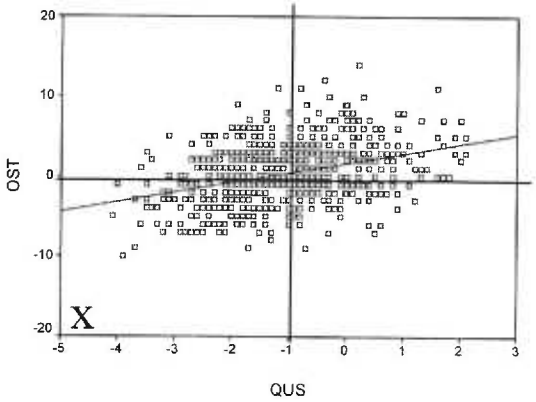

Table 4: Spearman correlations of comparisons of each predictor.

Comparison	Spearman Coefficient
QUS T ≤ -1 vs. ORAI	-0.382
QUS T ≤ -1 vs. OST	0.361
QUS T ≤ -1 vs. SCORE	-0.402
ORAI vs. OST	-0.909
ORAI vs. SCORE	0.868
OST vs. SCORE	-0.946

All coefficients statistically significant at 0.01 level.

Predictive Measures

Comparisons of each predictive scheme's predictive and selective abilities were made at each endpoint. The results for each endpoint are described in the tables and discussion below.

Femoral Neck

Femoral Neck T-Score ≤ -2

GEE logistic models were used to compare the difference between sensitivity, specificity, PPV, and NPV among the four predictive schemes at each endpoint. The results at this endpoint are shown in table 5, and the comparisons are detailed in table 6. At the femoral neck T-score ≤ -2 cutoff, the OST index demonstrates a significantly lower sensitivity

and a significantly higher specificity than each of the other three indices. While no other comparisons of sensitivity are significant at this endpoint, the specificity of QUS proves to be significantly higher than that of either the SCORE or ORAI. The ORAI also proves to have a significantly higher specificity than the SCORE index. Similarly, with respect to PPV and NPV, the OST index has a higher PPV and a lower NPV than each of the other pedictors. The comparisons between OST and QUS, however, do not achieve statistical significance for either PPV or NPV. Additionally, the ORAI index and QUS testing both have a significantly better PPVs than the SCORE index.

Table 5: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of femoral neck DXA T-Score $\leq$ -2. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T $\leq$ -1	0.878*	0.546 (0.502-0.590)	0.358 (0.309-0.410)	0.939 (0.906-0.964)	0.802 (0.764-0.841)
SCORE	0.959 (0.913-0.985)	0.350 (0.309-0.394)	0.298 (0.257-0.342)	0.967 (0.930-0.988)	0.788 (0.749-0.827)
ORAI	0.912 (0.854-0.952)	0.438 (0.395-0.483)	0.319 (0.275-0.366)	0.945 (0.908-0.970)	0.756 (0.713-0.800)
OST	0.714 (0.634-0.786)	0.662 (0.620-0.703)	0.379 (0.322-0.439)	0.889 (0.853-0.919)	0.776 (0.736-0.815)

* QUS T-score threshold set at $\leq$ -1 post hoc to obtain sensitivity within range of other predictors

Table 6: GEE Statistical Comparison of Differences in Predictive Measures at FN T $\leq$ -2

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T $\leq$ -1 vs. ORAI	--	--	13.95	<0.005‡	6.57	0.01	0.11	0.74
QUS T $\leq$ -1 vs. OST	--	--	16.05	<0.005‡	1.06	0.30	6.26	0.01
QUS T $\leq$ -1 vs. SCORE	--	--	46.93	<0.005‡	17.97	<0.005‡	2.27	0.13
ORAI vs. OST	29.40	<0.005‡	128.02	<0.005‡	16.63	<0.005‡	9.93	<0.005‡
ORAI vs. SCORE	5.31	0.02	22.70	<0.005‡	5.94	<0.005‡	2.59	0.11
OST vs. SCORE	30.91	<0.005‡	199.74	<0.005‡	25.60	<0.005‡	11.40	<0.005‡

‡ Statistically significant difference at p $\leq$ 0.0083 (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

Examining the areas under the ROC curves for each predictor at a femoral neck cutoff of -2, Figure 8 shows that the ROC curves are fairly closely matched. The ORAI seems to encompass a slightly smaller area than the other three, and indeed it does show the smallest AUROC value in table 5. Table 7 shows that AUROC for the ORAI index is significantly smaller than both the SCORE and the OST indices. None of the other differences in AUROC values between predictors is statistically significant.

Figure 8: AUROC Plot at Endpoint of FN T ≤ -2

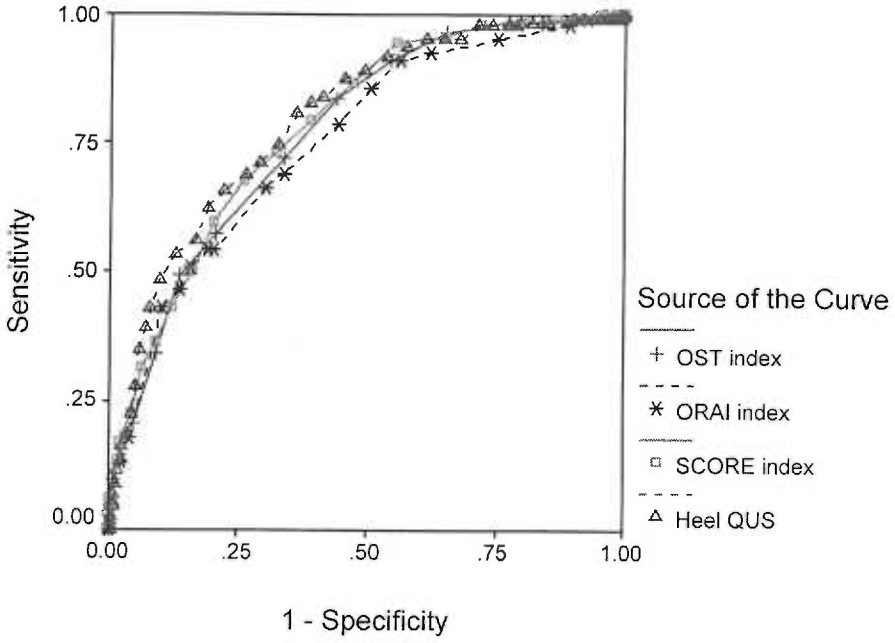

Table 7: Statistical Comparison of AUROC Differences at FN T ≤ -2

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.046	-	0.0706
QUS T ≤ -1 vs. OST	0.027	0.2829	0.2790
QUS T ≤ -1 vs. SCORE	0.013	-	0.5929
ORAI vs. OST	-0.019	-	0.0476**
ORAI vs. SCORE	-0.032	0.0284**	0.0122**
OST vs. SCORE	-0.011	-	0.1977

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at p ≤ 0.05

Femoral Neck T-Score ≤ -2.5

Similar to the results at FN ≤ -2, the specificity is significantly different for comparisons between each predictor, with the OST scoring highest, followed by QUS, the ORAI, and SCORE indices respectively. However, while comparisons of sensitivity again rate the OST lower than all three of the other predictors, only the SCORE vs. OST and ORAI vs. OST comparisons prove to be statistically significant. There are no significant differences among NPV for any of the predictors at this cutoff, but the PPV of the SCORE index proves significantly lower than all three other predictors, and the OST index has a significantly better PPV than either the SCORE or ORAI indices.

Table 8: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of femoral neck DXA T-Score ≤ -2.5. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T ≤ -1	0.929*	0.497 (0.455-0.538)	0.181 (0.142-0.224)	0.983 (0.961-0.995)	0.852 (0.809-0.895)
SCORE	0.971 (0.899-0.997)	0.311 (0.274-0.350)	0.143 (0.112-0.178)	0.989 (0.961-0.979)	0.831 (0.782-0.880)
ORAI	0.971 (0.901-0.997)	0.399 (0.359-0.440)	0.162 (0.128-0.201)	0.992 (0.968-0.999)	0.817 (0.769-0.864)
OST	0.814 (0.703-0.897)	0.625 (0.584-0.664)	0.206 (0.160-0.258)	0.966 (0.942-0.982)	0.822 (0.774-0.871)

* QUS T-score threshold set at ≤ -1 post hoc to obtain sensitivity within range of other predictors

Table 9: GEE Statistical Comparison of Differences in Predictive Measures at FN T ≤ -2.5

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T ≤ -1 vs. ORAI	--	--	13.95	<0.005‡	4.39	0.04	0.69	0.41
QUS T ≤ -1 vs. OST	--	--	22.93	<0.005‡	3.96	0.05	1.65	0.20
QUS T ≤ -1 vs. SCORE	--	--	52.72	<0.005‡	19.82	<0.005‡	0.29	0.59
ORAI vs. OST	8.47	<0.005‡	149.89	<0.005‡	21.57	<0.005‡	4.30	0.04
ORAI vs. SCORE	0.00	1.00	28.33	<0.005‡	16.39	<0.005‡	0.12	0.73
OST vs. SCORE	8.47	<0.005‡	231.32	<0.005‡	41.96	<0.005‡	2.86	0.09

‡ Statistically significant difference at p ≤ 0.0083. (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

AUROC comparison at FN T ≤ -2.5 looks even more closely aligned than at ≤-2, with no statistically significant differences in the comparisons of the areas listed in table 10, and a very tight visual representation of the plotted ROC curves (Fig. 9).

Figure 9: AUROC Plot at Endpoint of FN T ≤ -2.5

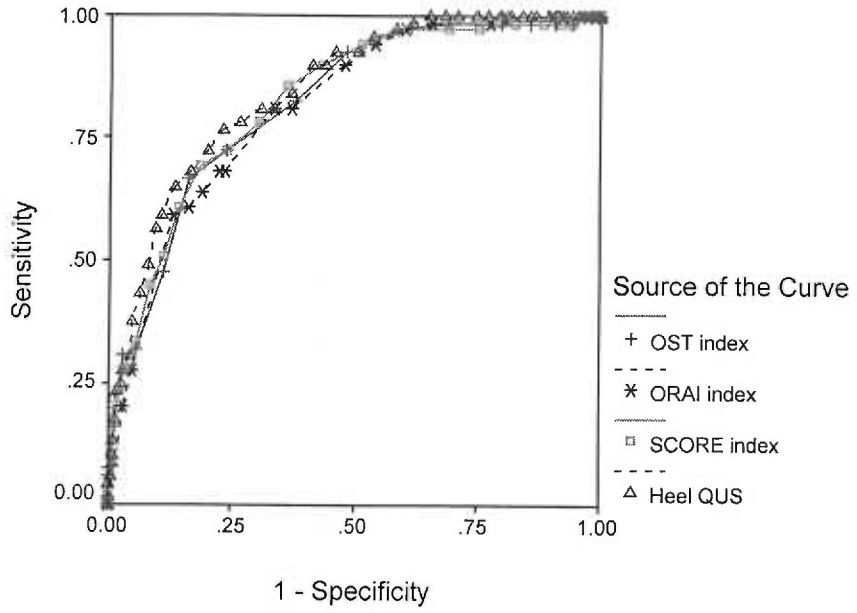

Table 10: Statistical Comparison of AUROC Differences at FN T ≤ -2.5

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.035	-	0.2130
QUS T ≤ -1 vs. OST	0.029	0.3172	0.3148
QUS T ≤ -1 vs. SCORE	0.019	-	0.5210
ORAI vs. OST	-0.006	-	0.6065
ORAI vs. SCORE	-0.013	0.4497	0.3940
OST vs. SCORE	-0.005	-	0.6363

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at p ≤ 0.05

Lumbar Spine

Lumbar Spine T-Score ≤ -2

Once again, the OST index has a significantly lower sensitivity and significantly higher specificity than all three of the other predictors. QUS has a significantly higher specificity than the remaining two indices, and the ORAI index's specificity is significantly better than that of the SCORE. None of the comparisons of NPV are significant at this endpoint, but the OST index and QUS do have a significant higher PPV than the SCORE index, and the OST is significantly higher than the ORAI as well.

Table 11: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of lumbar spine DXA T-Score ≤ -2. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T ≤ -1	0.813*	0.551 (0.506-0.597)	0.409 (0.358-0.461)	0.886 (0.844-0.919)	0.770 (0.730-0.811)
SCORE	0.890 (0.835-0.931)	0.345 (0.302-0.389)	0.340 (0.298-0.385)	0.891 (0.837-0.932)	0.718 (0.675-0.761)
ORAI	0.835 (0.773-0.886)	0.432 (0.387-0.478)	0.359 (0.314-0.407)	0.873 (0.824-0.913)	0.704 (0.661-0.748)
OST	0.648 (0.574-0.718)	0.660 (0.616-0.703)	0.421 (0.363-0.482)	0.831 (0.790-0.868)	0.704 (0.660-0.747)

* QUS T-score threshold set at ≤ -1 post hoc to obtain sensitivity within range of other predictors

Table 12: GEE Statistical Comparison of Differences in Predictive Measures at LS T ≤ -2

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T ≤ -1 vs. ORAI	--	--	17.08	<0.005‡	7.84	0.01	0.27	0.60
QUS T ≤ -1 vs. OST	--	--	13.93	<0.005‡	0.36	0.55	5.48	0.02
QUS T ≤ -1 vs. SCORE	--	--	51.96	<0.005‡	17.14	<0.005‡	0.06	0.81
ORAI vs. OST	34.47	<0.005‡	125.16	<0.005‡	15.79	<0.005‡	4.90	0.03
ORAI Vs. SCORE	5.60	0.02	21.98	<0.005‡	3.51	0.06	1.03	0.31
OST Vs. SCORE	47.12	<0.005‡	190.02	<0.005‡	22.67	<0.005‡	6.68	0.01

‡ Statistically significant difference at p ≤ 0.0083. (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

Comparing the ROC curves at this endpoint, QUS testing does appear to deviate to the higher side of the curve, indicating a larger AUROC than the other predictors, which appear similar to one another. QUS does prove to have a larger AUROC, and the difference is statistically significant when QUS is compared to the ORAI or OST indices.

Figure 10: AUROC Plot at Endpoint of LS T ≤ -2

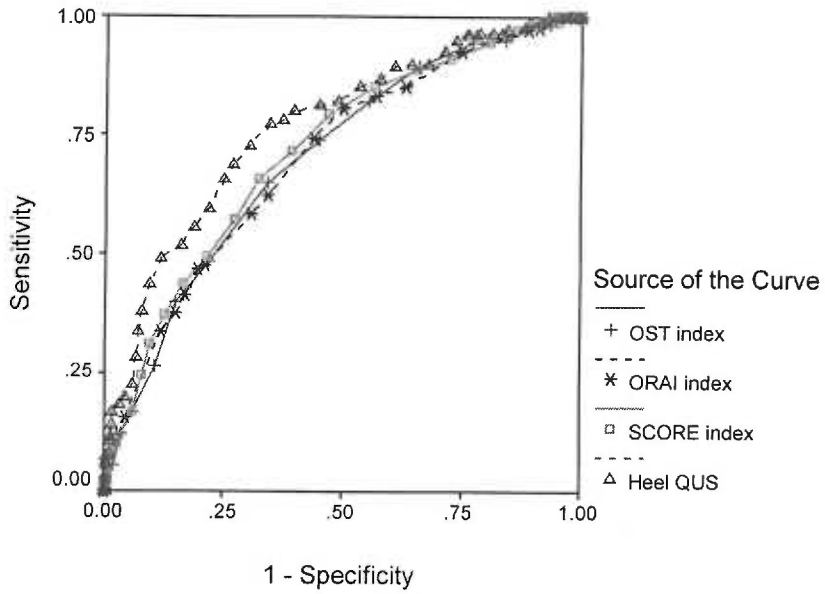

Table 13: Statistical Comparison of AUROC Differences at LS T ≤ -2

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.066	-	0.0139**
QUS T ≤ -1 vs. OST	0.066	0.0124**	0.0132**
QUS T ≤ -1 vs. SCORE	0.050	-	0.0578
ORAI vs. OST	0.001	-	0.9495
ORAI vs. SCORE	-0.014	0.3210	0.2582
OST vs. SCORE	-0.014	-	0.0891

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at $p \leq 0.05$

Lumbar Spine T-Score ≤ -2.5

The sensitivity and specificity at the endpoint of LS T ≤ –2.5 follow exactly the same pattern as at -2. The only statistically significant comparisons of sensitivity are those between the OST and each of the other predictors. In each of these cases, the OST has a significantly lower sensitivity than the other predictor. All comparisons of specificity are once again significant, with the OST having the highest specificity, followed by QUS, the ORAI and SCORE indices respectively. PPV and NPV comparisons are also similar to those at the previous endpoint, but in addition to the three significant PPV comparisons at the -2 cutoff, which are again significant at -2.5, the PPV of QUS is significantly higher than the PPV of the ORAI, and the NPV of the OST index is significantly higher than that of QUS.

Table 14: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of lumbar spine DXA T-Score ≤ -2.5. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T ≤ -1	0.895*	0.523 (0.480-0.566)	0.282 (0.236-0.331)	0.960 (0.931-0.979)	0.798 (0.755-0.841)
SCORE	0.903 (0.833-0.950)	0.318 (0.279-0.359)	0.216 (0.179-0.256)	0.940 (0.896-0.970)	0.731 (0.681-0.780)
ORAI	0.868 (0.792-0.924)	0.406 (0.364-0.448)	0.234 (0.195-0.277)	0.936 (0.897-0.964)	0.726 (0.676-0.775)
OST	0.684 (0.591-0.768)	0.629 (0.587-0.670)	0.279 (0.227-0.335)	0.905 (0.871-0.933)	0.716 (0.666-0.766)

* QUS T-score threshold set at ≤ -1 post hoc to obtain sensitivity within range of other predictors

Table 15: GEE Statistical Comparison of Differences in Predictive Measures at LS T ≤ -2.5

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T ≤ -1 vs. ORAI	--	--	18.71	<0.005‡	13.14	<0.005‡	1.90	0.17
QUS T ≤ -1 vs. OST	--	--	14.76	<0.005‡	0.00	0.95	9.11	<0.005‡
QUS T ≤ -1 vs. SCORE	--	--	58.99	<0.005‡	28.99	<0.005‡	1.12	0.29
ORAI vs. OST	19.52	<0.005‡	140.56	<0.005‡	12.39	<0.005‡	3.78	0.05
ORAI vs. SCORE	1.32	0.25	26.92	<0.005‡	4.77	0.03	0.06	0.80
OST vs. SCORE	25.57	<0.005‡	213.45	<0.005‡	21.85	<0.005‡	3.47	0.06

‡ Statistically significant difference at p ≤ 0.0083. (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

The difference between the ROC curve of QUS and that of the other predictors is even more pronounced at the lumbar spine T-score cutoff of -2.5 (Fig.11). At this threshold, the AUROC difference is significant between QUS and each of the other predictors.

Figure 11: AUROC Plot at Endpoint of LS T ≤ -2.5

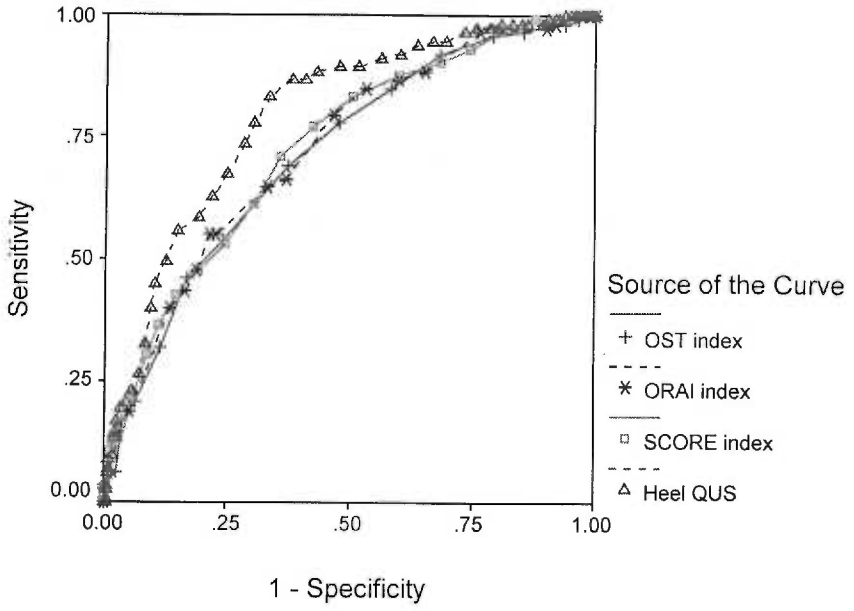

Table 16: Statistical Comparison of AUROC Differences at LS T ≤ -2.5

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.072	-	0.0118**
QUS T ≤ -1 vs. OST	0.082	0.0041**	0.0040**
QUS T ≤ -1 vs. SCORE	0.065	-	0.0272**
ORAI vs. OST	0.010	-	0.4231
ORAI vs. SCORE	-0.005	0.7496	0.7328
OST vs. SCORE	-0.014	-	0.1750

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at p ≤ 0.05

Either Femoral Neck or Lumbar Spine Values Below Threshold

Femoral Neck or Lumbar Spine T-Score ≤ -2

Examining the comparisons of sensitivity and specificity at this endpoint, the same pattern seen previously with the OST index, of a significantly lower sensitivity and higher specificity than the other indices, holds true. In addition to these comparisons, QUS demonstrates a higher specificity than the ORAI and SCORE indices, respectively, which are also significantly different. Additionally, the SCORE index does have a significantly higher sensitivity than the ORAI. The OST demonstrates a significantly lower NPV than each of the other indices, and a significantly higher PPV than the ORAI and SCORE indices. QUS testing is also significantly higher in PPV than the ORAI and SCORE.

Table 17: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of femoral neck or lumbar spine DXA T-Score ≤ -2. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T ≤ -1	0.827*	0.598 (0.550-0.645)	0.521 (0.468-0.574)	0.868 (0.824-0.904)	0.800 (0.764-0.836)
SCORE	0.902 (0.856-0.938)	0.379 (0.333-0.427)	0.434 (0.388-0.480)	0.880 (0.825-0.924)	0.760 (0.722-0.799)
ORAI	0.845 (0.791-0.890)	0.470 (0.423-0.518)	0.457 (0.408-0.506)	0.852 (0.800-0.895)	0.738 (0.698-0.778)
OST	0.655 (0.589-0.717)	0.703 (0.658-0.746)	0.538 (0.477-0.598)	0.794 (0.750-0.834)	0.746 (0.706-0.785)

* QUS T-score threshold set at ≤ -1 post hoc to obtain sensitivity within range of other predictors

Table 18: GEE Statistical Comparison of Differences in Predictive Measures at FN or LS ≤ -2

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T ≤ -1 vs. ORAI	--	--	16.58	<0.005‡	10.36	<0.005‡	0.41	0.52
QUS T ≤ -1 vs. OST	--	--	11.39	<0.005‡	0.50	0.48	8.82	<0.005‡
QUS T≤ -1 vs. SCORE	--	--	48.49	<0.005‡	21.97	<0.005‡	0.23	0.63
ORAI vs. OST	44.43	<0.005‡	114.24	<0.005‡	21.14	<0.005‡	8.10	<0.005‡
ORAI vs. SCORE	8.12	<0.005‡	19.62	<0.005‡	4.33	0.04	2.17	0.14
OST vs. SCORE	59.27	<0.005‡	175.08	<0.005‡	29.62	<0.005‡	11.45	<0.005‡

‡ Statistically significant difference at p ≤ 0.0083. (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

This is the most inclusive of the 6 endpoints, as cases with T-scores at or below -2 for either the femoral neck or lumbar spine are considered as part of the outcome for this endpoint. Observation of the plotted ROC curves shows that QUS still appears to have a slightly greater area than the other predictors, but that the difference seen in the lumbar spine ROC graphs seems to have been diluted somewhat. The area for QUS was significantly higher in comparison to the ORAI and OST indices, but not to the SCORE index.

Figure 12: AUROC Plot at Endpoint of LS or FN T ≤ -2

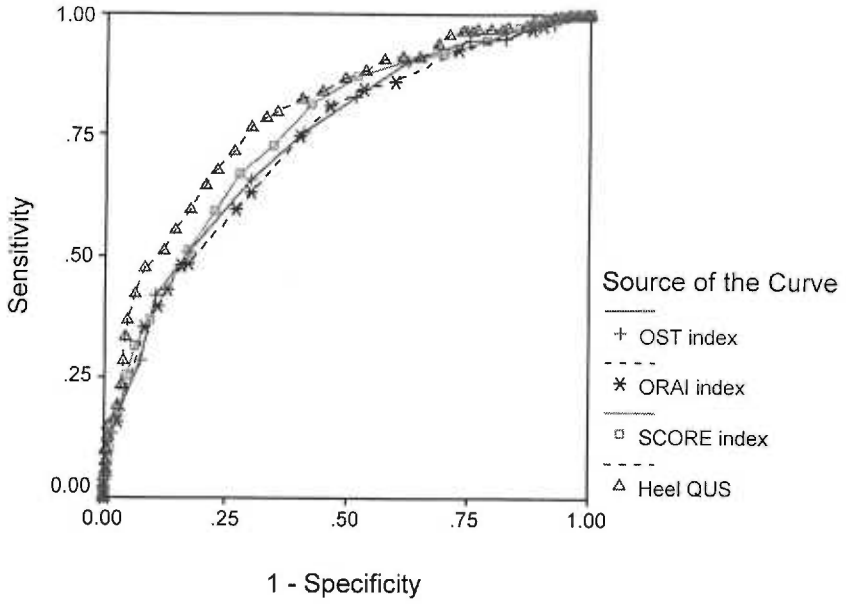

Table 19: Statistical Comparison of AUROC Differences at FN or LS ≤ -2

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.062	-	0.0111**
QUS T ≤ -1 vs. OST	0.054	0.0262**	0.0257**
QUS T ≤ -1 vs. SCORE	0.039	-	0.1083
ORAI vs. OST	-0.008	-	0.4394
ORAI vs. SCORE	-0.023	0.0986	0.0608
OST vs. SCORE	-0.014	-	0.0667

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at p ≤ 0.05

Femoral Neck or Spine T-Score ≤ -2.5

Lastly, for this endpoint specificity and sensitivity look just as they do for almost all the other endpoints: the OST has significantly greater specificity and lower sensitivity, while QUS has a significantly higher specificity than the ORAI or SCORE. In PPV comparisons, both QUS and the OST prove significantly higher than the SCORE and ORAI indices. For NPV, only the comparison between OST and QUS is significant, with the QUS value higher.

Table 20: Summary of sensitivity, specificity, PPV, and NPV of the four predictive schemes at an endpoint of femoral neck or lumbar spine DXA T-Score ≤ -2.5. 95% confidence intervals are in parentheses.

Predictor	Sensitivity	Specificity	PPV	NPV	AUROC
QUS T ≤ -1	0.884*	0.541 (0.497-0.584)	0.340 (0.291-0.391)	0.946 (0.913-0.969)	0.818 (0.779-0.856)
SCORE	0.912 (0.852-0.954)	0.334 (0.293-0.377)	0.267 (0.228-0.310)	0.935 (0.889-0.966)	0.765 (0.720-0.809)
ORAI	0.877 (0.810-0.927)	0.424 (0.381-0.468)	0.289 (0.246-0.336)	0.928 (0.887-0.958)	0.752 (0.707-0.797)
OST	0.696 (0.612-0.771)	0.653 (0610.-0.694)	0.349 (0.293-0.409)	0.889 (0.853-0.919)	0.749 (0.703-0.794)

* QUS T-score threshold set at ≤ -1 post hoc to obtain sensitivity within range of other predictors

Table 21: GEE Statistical Comparison of Differences in Predictive Measures at FN or LS ≤ -2.5

	Sensitivity		Specificity		PPV		NPV	
Comparison	F***	p-value	F***	p-value	F***	p-value	F***	p-value
QUS T ≤ -1 vs. ORAI	--	--	17.04	<0.005‡	11.15	<0.005‡	0.90	0.34
QUS T ≤ -1 vs. OST	--	--	15.44	<0.005‡	0.27	0.60	8.30	<0.005‡
QUS T ≤ -1 vs. SCORE	--	--	55.54	<0.005‡	27.05	<0.005‡	0.29	0.59
ORAI vs. OST	23.73	<0.005‡	136.05	<0.005‡	17.59	<0.005‡	5.13	0.02
ORAI vs. SCORE	1.91	0.17	26.12	<0.005‡	6.28	0.01	0.19	0.67
OST vs. SCORE	30.43	<0.005‡	208.13	<0.005‡	29.29	<0.005‡	5.05	0.03

‡ Statistically significant difference at p ≤ 0.0083. (Bonferroni adjustment for multiple comparisons)
*** Wald F statistic; d.f. for all comparisons is 1.
Comparisons of sensitivity involving QUS not evaluated due to post hoc nature of threshold determination.

Finally, the ROC plot seen in figure 13 also shows a curve for QUS that seems to encompass more area than the others. Again, the difference is not as pronounced as that for the lumbar spine alone, but it is visible nonetheless. The SCORE index also appears to diverge from the other predictors slightly, and does show a higher AUROC than the OST or ORAI in table 20. This ordering is also illustrated by the fact that the AUROC for QUS is significantly higher than that for the ORAI and OST, but not for the SCORE.

Figure 13: AUROC Plot at Endpoint of LS or FN T ≤ -2.5

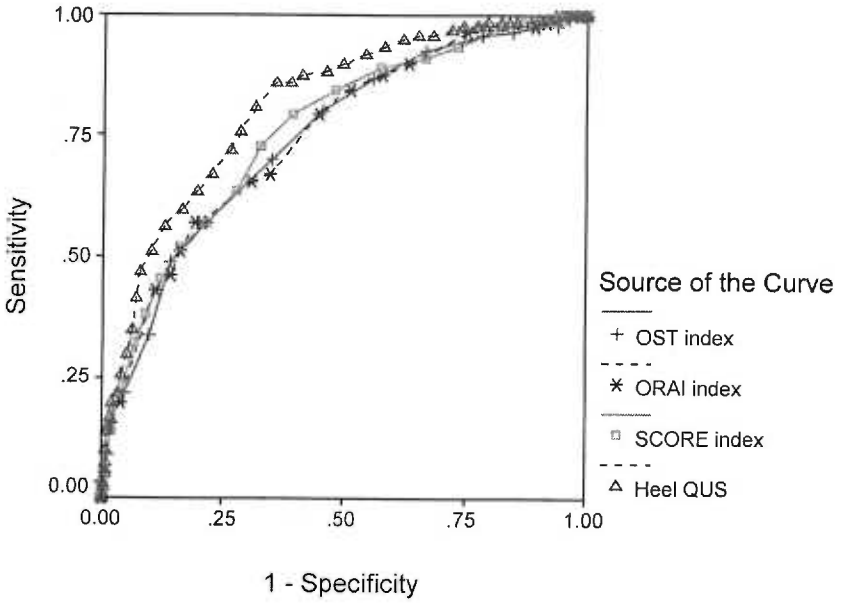

Table 22: Statistical Comparison of AUROC Differences at FN or LS ≤ -2.5

Comparison	Diff	p-value*	p-value†
QUS T ≤ -1 vs. ORAI	0.066	-	0.0127**
QUS T ≤ -1 vs. OST	0.069	0.0089**	0.0081**
QUS T ≤ -1 vs. SCORE	0.051	-	0.0551
ORAI vs. OST	0.004	-	0.7450
ORAI vs. SCORE	-0.013	0.4079	0.3517
OST vs. SCORE	-0.015	-	0.0928

*Using method of Hanley and McNeil (1983)
†Using method of Delong, Delong, and Clarke-Pearson (1988)
**Statistically significant difference at p ≤ 0.05

Summary of Results

In the analysis of specificity and sensitivity, the OST index was the only predictor that consistently performed significantly differently from the other predictors in both measurements. Its sensitivity was the lowest and its specificity the highest across all 6 endpoints. In GEE testing, the differences in specificity between the OST and each of the other predictors proved significant for all endpoints tested in this manner, and the differences in sensitivity were significant against all 3 clinical indices. No other tests of

the comparison of sensitivity between individual predictors proved to be significant except that of the SCORE and ORAI at FN or LS ≤-2. All GEE tests of individual comparisons of specificity, however, proved to be significant at all endpoints. Again, the OST had the highest specificity (offset by its significantly lower sensitivity in all but one comparison), followed by QUS testing, then by the ORAI, and the SCORE had the lowest specificity. This ordering held true for all 6 endpoints, and the differences between individual specificities were significant for all of these comparisons.

Due to its independence from any assigned threshold cutoff, and its effective combination of sensitivity and specificity measures, AUROC testing is of particular utility in evaluating diagnostic tests, and in this case comparing the 4 predictive schemes. For all 6 endpoints, QUS testing had the highest AUROC score, followed by the SCORE index. The significance of the comparisons, however, was variable. In the femoral neck, the AUROC for ORAI was significantly lower than that of the SCORE and OST indices at a T-Score of -2. No other AUROC comparisons were significant at either femoral neck threshold, however. At the lumbar spine, QUS was significantly better then the OST and the ORAI at a T-score of -2, and significantly better than all 3 other predictors at a T-score of -2.5. At the endpoints of T-scores ≤ -2 and -2.5 at either LS or FN, QUS remained significantly better than the ORAI and OST at both cutoffs.

E. Discussion

The results of the analysis demonstrate that at the femoral neck, none of the predictive schemes are clearly or significantly better at predicting low bone density than the others, while at the lumbar spine QUS testing performs slightly better overall than the clinical

indices. This slightly better performance in QUS carries over when DXA thresholds are examined at either the femoral neck or lumbar spine sites. Observing the ROC curves at each endpoint, it is difficult to differentiate among the predictors for each of the femoral neck endpoints (Figs. 8,9), while at the lumbar spine (Figs. 10,11) QUS clearly encompasses more area on each of the graphs. The differences are still apparent at the endpoints including either femoral neck or lumbar spine (Figs. 12,13), but the effect appears to be somewhat watered down.

In evaluating the outcomes of performance measures other than AUROC, it is useful by way of illustration to calculate how many patients in this study who did have low DXA bone density at a given threshold would be identified for referral by each of the predictive schemes (corresponding to the predictor's sensitivity) as well as how many patients not having bone density at or below the threshold would be indicated for referral (corresponding to 1 – specificity). Using hip or spine T-scores at or below -2, which is the most inclusive DXA endpoint in the study and is also recommended by the North American Menopause Society in its treatment recommendations,[29, 30] table 23 illustrates the number of patients in each of these categories for each predictor. As the table shows, out of the 226 patients whose DXA scores at the femoral neck or lumbar spine were below the endpoint of $T \leq -2$, the SCORE index would have chosen 204 of them for referral, while the OST index would have chosen only 148 patients. However, the SCORE index also would refer 266 women whose DXA T-score was above that threshold, while the OST index would identify only 121 of those subjects for referral. The results in table 23 are dependent on the cutoffs used for each scoring system, but

reflect the balance between selecting patients who are good candidates for DXA testing and selecting those for whom it is likely unnecessary.

Table 23: Number of patients correctly and incorrectly chosen for DXA referral by predictive schemes, using three separate endpoints.

	FN or LS T-score ≤ -2	
Predictor	Selected and below DXA threshold	Selected but not below DXA threshold
QUS	187/226*	172/428†
SCORE	204/226*	266/428†
ORAI	191/226*	227/428†
OST	148/226*	121/428†

* second number represents number of patients with T-score less than or equal to -2 at given site.

† second number represents number of patients above T-score of -2 at given site.

Following is a discussion of the results for each of the four predictors.

The SCORE index

The SCORE index is the oldest of the predictive indices, and has had the most validation performed on it of any of the three. It uses age, any prior ERT use, fracture history, race, weight, and history of rheumatoid arthritis to predict the likelihood of low bone density. The range of and specificity found in this study for the SCORE index at the various endpoints is well within the range seen in other population-based studies validating the SCORE[11-19]. This may suggest that this study was not greatly affected by the referral-based nature of its population; however, if the study in which the SCORE questionnaire was initially developed[11] and those using a modified threshold cutoff score to produce a sensitivity of approximately 90%[18,19] are removed, the specificities seen at the various endpoints in this study (0.311-0.379) are slightly higher than those reported in the remaining studies. This could indeed reflect the previous screening that this population received, in the form of physician referral.

Table 24: Results frm other studies of the SCORE index performance

Study	DXA T-score Endpoint	Sensitivity	Specificity	AUROC
This study	FN ≤ -2	0.959	0.350	0.788
	FN ≤ -2.5	0.971	0.311	0.831
	LS ≤ -2	0.890	0.345	0.718
	LS ≤ -2.5	0.903	0.318	0.731
	Either LS or FN ≤ -2	0.902	0.379	0.760
	Either LS or FN ≤ -2.5	0.912	0.334	0.765
Lydick et al 1998[11]	FN ≤ -2	0.894	0.497	0.75
Cadarette SM et al 2001[12]	FN ≤ -2	0.975	0.208	0.77
	FN ≤ -2.5	0.996	0.179	0.80
Russell & Morrison 2001[13]	FN ≤ -2.5	0.994	0.271	--
	LS ≤ -2.5	0.960	0.280	
Von Muhlen D et al 1999[15]	FN ≤ -2	0.980	0.125	0.696
Cadarette SM et al 1999[16]	LS ≤ -2	0.933	0.257	--
	FN ≤ -2	0.904	0.290	--
	Either LS or FN ≤ -2	0.900	0.320	0.710
Ben Sedrine W et al 2001[17]	FN ≤ -2	0.915	0.265	--
Hochberg et al 2001[18]	Either LS or FN ≤ -2	0.88*	0.36*	--
	Either LS or FN ≤ -2.5	0.93*	0.31*	--
Siris et al 2001[19]	FN ≤ -2	0.87†	0.55†	--
	FN ≤ -2.5	0.90**	0.59**	--
	LS ≤ -2	0.90‡	0.35‡	--
	LS ≤ -2.5	0.89***	0.40***	--

* Uses modified SCORE index threshold cutoff of >10
† Uses modified threshold cutoff of > 6
** Uses modified threshold cutoff of > 7
‡ Uses modified threshold cutoff of > 4
*** Uses modified threshold cutoff of >5

The AUROC values for the SCORE index were consistently the second highest of the four predictors, behind only QUS testing, at all 6 endpoints. While the differences by and large are not statistically significant, the fact that the QUS fails to be significantly larger than the SCORE index at endpoints where it is significantly higher than both the ORAI and OST indices indicates that the SCORE does at least perform comparably well. This difference is visible in figures 12 and 13, where the ROC curve of the SCORE index is discernable from the OST and ORAI indices even though it clearly does not encompass as much area as the QUS curve does.

The ORAI index

The ORAI index uses age, which is categorized into 3 strata, weight, and current estrogen use to calculate a score. It did not appear to perform as well overall as the SCORE index or QUS testing, though the differences are not pronounced and of variable direction and significance. The ORAI fell in between the other predictors in most comparisons of sensitivity, specificity, PPV, and NPV, while its AUROC was significantly lower than that of QUS testing at the lumbar spine and at either site, but not at the femoral neck alone. The ORAI is fairly simple to calculate, and relies only on age category, weight, and whether or not the patient is currently on estrogen.

Table 25: Results from other studies of the ORAI index performance

Study	DXA T-score Endpoint	Sensitivity	Specificity	AUROC
This study	FN ≤ -2	0.912	0.438	0.756
	FN ≤ -2.5	0.971	0.399	0.817
	LS ≤ -2	0.835	0.432	0.704
	LS ≤ -2.5	0.868	0.406	0.726
	Either LS or FN ≤ -2	0.845	0.470	0.738
	Either LS or FN ≤ -2.5	0.877	0.424	0.752
Cadarette SM et al 2000[20]	Either LS or FN ≤ -2	0.900*	0.451*	--
	Either LS or FN ≤ -2.5	0.970*	0.413*	--
	Either LS or FN ≤ -2	0.933†	0.464†	--
	Either LS or FN ≤ -2.5	0.944†	0.414†	--
Cadarette SM et al 2001[12]	FN ≤ -2	0.942	0.319	0.76
	FN ≤ -2.5	0.975	0.278	0.79
Hochberg et al 2001[18]	Either LS or FN ≤ -2	0.89**	0.29**	--
	Either LS or FN ≤ -2.5	0.92‡	0.26‡	--
Siris et al 2001[19]	FN ≤ -2	0.89**	0.39**	--
	FN ≤ -2.5	0.90‡	0.52‡	--
	LS ≤ -2	0.92***	0.24***	--
	LS ≤ -2.5	0.91**	0.36**	--

* Development Cohort
† Validation Cohort
** Uses modified threshold cutoff of > 5
‡ Uses modified threshold cutoff of > 8
*** Uses modified threshold cutoff of > 3

Comparing the results of this study with other studies which test the performance of the ORAI, it is again of note that while the overall ranges of sensitivity and specificity are

similar, the specificities seen in this study appear slightly larger. It is impossible to make any valid statistical comparisons between the results of these different studies, but again this may reflect the referral-based nature of the population in this study.

The OST Index

The OST index is the most simple of the three predictive indices, and requires the least information to calculate – only age and weight. In almost every comparison, it demonstrated a significantly higher specificity and lower sensitivity than each of the other predictors. The only exception to this was that at the femoral neck endpoint of $T \leq -2.5$, the OST did not show a significantly lower sensitivity than QUS, despite retaining a significantly higher specificity. The fairly consistent nature of these differences may reflect the need for a slightly more inclusive cutoff threshold to be used with this index to predict low bone density. For example, at the endpoint of either femoral neck or lumbar spine T-score ≤ -2, the recommended cutoff of ≤ -1 yields a sensitivity of 0.655 and a specificity of 0.703. Moving the threshold to ≤ 4 includes more patients in the group recommended for treatment, and would raise the sensitivity to 0.903 at the expense of lowering the specificity to 0.376. This would bring these measures more in line with those of the other predictive schemes. This study was originally developed and validated on a population made up exclusively of Asian women (in some studies it is referred to as the OSTA, or Osteoporosis Self-Assessment Tool for Asians), and while Caucasian and Asian populations are typically regarded as having similar bone density and fracture risk profiles, it may be of interest in future studies to assess if the different results with the recommended cutoff is based partially on the Caucasian ethnicity of this population. It is

apparent from looking at table 26 that the sensitivity and specificity of the OST are more in line with the diagnostic profile of the other indices in this study when the cutpoint is raised, as in the studies by Hochberg et al and Siris et al. The study by Siris is comprised of 82% Caucasian women, while the Hochberg study population is entirely Caucasian.

Table 26: Results from other studies of the OST index performance

Study	DXA T-score Endpoint	Sensitivity	Specificity	AUROC
This study	FN ≤ -2	0.714	0.662	0.776
	FN ≤ -2.5	0.814	0.625	0.822
	LS ≤ -2	0.648	0.660	0.704
	LS ≤ -2.5	0.684	0.629	0.716
	Either LS or FN ≤ -2	0.655	0.703	0.746
	Either LS or FN ≤ -2.5	0.696	0.653	0.749
Koh et al 2001[21]	FN ≤ -2.5	0.91	0.45	0.79
Hochberg et al 2001[18]	Either LS or FN ≤ -2	0.83†	0.52†	--
	Either LS or FN ≤ -2.5	0.89*	0.46*	--
Siris et al 2001[19]	FN ≤ -2	0.90†	0.47†	--
	FN ≤ -2.5	0.89*	0.52*	--
	LS ≤ -2	0.90**	0.32**	--
	LS ≤ -2.5	0.93**	0.30**	--

* Uses modified threshold cutoff of ≤ 1
† Uses modified threshold cutoff of ≤ 2
** Uses modified threshold cutoff of ≤ 3

QUS Testing

Testing by QUS of the heel performed similar to the other predictors in the femoral neck, and statistically better than all three indices at a lumbar spine cutoff of T ≤ -2.5. Examination of the ROC curves of the predictors at each endpoint clearly show that the QUS ROC curve encompasses slightly more area at the lumbar spine endpoints, and is similar to the other predictors at the spine endpoints. This may or may not be reflective of the fact that the calcaneus, the heel bone measured by QUS, is made up of a combination of cortical and trabecular bone that is more similar to vertebrae than to the femoral neck.

Correlations

While it is not surprising that the predictive indices correlate better with each other than with QUS testing (Figs. 2-7), it is interesting that the tightest correlation is seen between the OST, which is the simplest of the predictors, and the SCORE, which is the most complex and requires the most information to calculate a score. This may be due to the broad age categorizations used by the ORAI index, while the OST and SCORE indices use calculations which, by truncating to the first integer, effectively categorize age in 10-year intervals. The correlation between the OST and SCORE index may also partially be explained by the relative homogeneity of this study population, which might not use the SCORE index's numerous adjustments to discriminate as effectively as a more varied population.

Limitations

This study was designed to examine the predictive ability of several clinical diagnostic schemes in a group of women referred for bone density testing. The generalizability of the study is limited due to the referral-based nature of this population; however, this is precisely the population in which we are interested in potentially applying some form of diagnostic screening. That is, women whose physicians have already subjectively identified them as potential candidates for DXA screening are the population that the results of this study are intended to apply to.

Another factor limiting the generalizability of this study is that the source population for this project is entirely of Caucasian ethnicity. As mentioned previously, due to the fact that QUS testing relies on a reference population comprised and validated only among

postmenopausal Caucasian women, technicians were instructed to perform QUS scans only on Caucasian women. With few exceptions, this is in large part consistent with the ethnicity of patients referred for bone density testing at the Center; however, given the largely Caucasian population of the Portland metropolitan area where the clinic is located, this data set may not be reflective of the ethnic composition of other areas. Possible disparities with regard to health insurance coverage and health care access between whites and non-whites may exaggerate the under-representation of minorities in the Center's referral population. It is also of note that the OST index was designed on a population of Asian women, and has not been validated significantly on a non-Asian population. Another of the indices, the SCORE index, uses race (African-American vs. non-African American) as one of its scoring criteria. Studies have shown women of African descent to be at less risk for osteoporosis-associated fractures than other women[36, 37], and the scoring system makes a significant adjustment (+5 for non-African-American out of a 6 or greater threshold for referral) for this criterion. Lack of African American patients in the study population will not reduce the index's predictive strength on an individual patient basis, but overall could weaken its generalizability to populations where African-American women would be included.

A potential limitation is the use of a study population that effectively amounts to a convenience sample. DXA technologists were instructed to perform QUS testing in addition to the standard DXA measurements as time permitted for postmenopausal Caucasian women with no secondary causes of osteoporosis, who were scheduled to be scanned on one particular DXA instrument. In addition, the technologists were told to attempt to maintain an even distribution of age among women scanned, in order to ensure

a complete picture of how heel ultrasound could be used in the Center's referral population. This introduces the possibility of selection bias into this study, by using a study population which was generated by somewhat subjective criteria and the technologist's ability (and willingness) to scan patients as time permitted. However, even in a completely subjective selection scheme, there would be no way for the technologists to have knowingly chosen patients to receive QUS testing based on the QUS score, which they could not yet know, or on the patient's scores on the clinical indices, which will not be calculated until all data is extracted and no longer personally identifiable. While the technologists would know hip and spine DXA scores, they would have no way of knowing the status of the predictor variables examined in this study.

Another limitation of this study is the differing regions and thresholds for measuring low bone density by DXA. While the World Health Organization defines the disease of osteoporosis as a bone mineral density score of 2.5 or more standard deviations below the young normal value, there is some debate as to the appropriate use of this cutoff, and to the adequate threshold to designate a patient as a candidate for therapy. As discussed previously, the North American Menopause society advocates a threshold T-score of -2 in younger, otherwise normal postmenopausal women for consideration of treatment, as well as in all postmenopausal women with at least one risk factor for hip fracture.[29, 30] There are two common anatomical sites for DXA bone density measurement to evaluate risk of fracture, the lumbar spine (typically L1-L4) and the femoral neck. To address these varying measurements and sites, comparisons were made using both 2 and 2.5 standard deviations below young normal values as threshold definitions of low bone density. Additionally, the performance of the predictive schemes was evaluated in

predicting patients with DXA bone density below each respective threshold value at either the femoral neck or lumbar spine sites.

A further limitation of this study, although unlikely, is the possibility that since the paper describing the creation of the SCORE index was published over three years ago, some physicians are already employing such indices in their decisions about whom to refer or not refer for bone density. However, since one of the aims of this study is to demonstrate a difference between the performance of the indices and current physician practices in predicting low bone density, any bias resulting from physicians using the indices being evaluated instead of their own subjective judgement would bias the results towards demonstrating no difference between physician referral and index performance.

F. Summary and Conclusions

The four predictive schemes evaluated in this study are similar in their ability to predict low bone density in this population, with QUS testing performing modestly better than the other predictors in diagnosing low bone density at the lumbar spine. Use of any one of these methods by primary care physicians could significantly reduce the number of women unnecessarily referred for bone density testing, at a potentially large cost savings to patients and health care organizations.

Table 27: Positive Predictive Value of Physician Referral Alone in the Study Population

DXA T-score Endpoint	Number of Patients	Number at or Below Threshold	PPV
FN $\leq$ -2	656	147	0.224
FN $\leq$ -2.5	656	70	0.107
LS $\leq$ -2	659	182	0.276
LS $\leq$ -2.5	659	114	0.173
FN or LS $\leq$ -2	654	226	0.346
FN or LS $\leq$ -2.5	654	138	0.211

While a direct comparison is not useful or statistically valid due to the fact that physician screening in the form of referral for DXA has already occurred in this population, it is interesting to note that all 4 predictive schemes had a higher PPV at each endpoint than those shown for physician referral in table 27. While, again, direct comparison between the PPVs in this table and those of the predictive schemes evaluated in this study would be inappropriate, it is evident that applying any of the predictive schemes to patients whom physicians had already decided were candidates for DXA testing could significantly reduce the number of patients unnecessarily made to undergo and pay for the procedure.

QUS testing, while performing modestly better than the clinical indices in predicting women with low bone density at the lumbar spine, still represents an additional procedure for patients to undergo, whether they are subsequently referred for DXA or not. The associated cost, time, and effort required of the patient and provider makes the clinical utility of screening for DXA using QUS dubious, given that the clinical indices do not perform significantly different at the femoral neck, and predict a low DXA value only slightly less well than QUS at the lumbar spine. These indices also have the advantage of being free, are based on clinical information that is likely already collected at the visit, and do not require the patient to undergo an additional procedure. Of the indices, the OST index requires only age and weight for calculation, and its AUROC values are not significantly different than the other clinical indices except at the femoral neck, where it performs significantly better than the ORAI. Based on the results of its sensitivity and specificity using the cutpoint employed in this study, however, it may be useful in populations similar to this one to use a different and more inclusive threshold value to

determine candidates for DXA testing.

G. Literature Cited

1. National Institutes of Health. *NIH Consensus Statement, Osteoporosis Prevention, Diagnosis, and Therapy.*17(1) March 27 –29,2000

2. WHO Study Group. *Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: Report of a WHO Study Group.* WHO Technical Report Series 843, Geneva, Switzerland: World Health Organization, 1994:1-129

3. Gehlbach SH et al. Recognition of osteoporosis by primary care physicians. American Journal of Public Health 2002; 92: 271-273

4. Abbott TA et al. Efficient patient identification strategies for women with osteoporosis. Journal of Clinical Densitometry 1999; 2(3):223-230

5. Van der Voot DJM et al. Screening for osteoporosis using easily obtainable clinical biometrical data: Diagnostic accuracy of measured, self-reported and recalled BMI, and related costs of bone mineral density measurements. Osteoporosis International 2000; 11:233-239

6. Kroger H, Reeve J. Diagnosis of osteoporosis in clinical practice. Ann Med 1998; 30: 278-287

7. Goemaere S et al. Limited clinical utility of a self-evaluating risk assessment scale for postmenopausal osteoporosis: Lack of predictive value of lifestyle-related factors. Calcified Tissue International 1999; 65: 354-358

8. Blum M et al. Weight and body mass index at menarche are associated with premenopausal bone mass. Osteoporosis International 2001; 12: 588-594

9. Ryan PJ et al. Postmenopausal screening for osteopenia. British Journal of Rheumatology 1992; 31:823-828

10. Versluis, RGJA et al. Clinical risk factors as predictors of postmenopausal osteoporosis in general practice. British Journal of General Practice 2001; 51: 806-810

11. Lydick et al. Development and validation of a simple questionnaire to facilitate identification of women likely to have low bone density. American Journal of Managed Care 1998; 4: 37-48

12. Cadarette SM et al. Evaluation of Decision Rules for Referring Women for Bone Densitometry by Dual-Energy X-Ray Absorptiometry. JAMA 2001; 286(1):57-63

13. Russell AS, Morrison RT. An assessment of the new "SCORE" index as a predictor of osteoporosis in women. Scandinavian Journal of Rheumatology 2001; 30: 35-39

14. Ungar WJ et al. The Canadian SCORE questionnaire. Journal of Clinical Densitometry 2000; 3(3):269-280

15. Von Muhlen D et al. Evaluation of the Simple Calculated Osteoporosis Risk Estimation (SCORE) in older Caucasian women: The Rancho Bernardo study. Osteoporosis International 1999; 10: 79-84

16. Cadarette SM et al. Validation of the Simple Calculated Osteoporosis Risk Estimation (SCORE) for patient selection for bone densitometry. Osteoporosis International 1999; 10: 85-90

17. Ben Sedrine W et al. Evaluation of the Simple Calculated Osteoporosis Risk Estimation (SCORE) in a sample of white women from Belgium. Bone 2001; 29(4): 374-380

18. Hochberg MC et al. Testing and Validation of Simple Clinical Risk Indices to Identify Postmenopausal Women with Osteoporosis. Journal of Bone and Mineral Research 2001; 16(Supp 1) S277 [abstract].

19. Siris E et al. Performance of Self-Assessment Risk Indices for Encouraging Appropriate Use of Bone Density Measurements among Postmenopausal Women. Journal of Bone and Mineral Research 2001; 16(Supp 1) S341 [abstract].

20. Cadarette SM et al. Development and validation of the Osteoporosis Risk Assessment Instrument to facilitate selection of women for bone densitometry. Canadian Medical Association Journal 2000; 162(9):1289-1294

21. Koh LKH et al. A Simple Tool to Identify Asian Women at Increased Risk of Osteoporosis. Osteoporosis International 2001; 12(8):699-705.

22. Yeap SS et al. The relationship between bone mineral density and ultrasound in postmenopausal and osteoporotic women. Osteoporosis International 1998 8:141-146

23. Frost ML et al. Quantitative ultrasound and bone mineral density are equally strongly associated with risk factors for osteoporosis. Journal of Bone and Mineral Research 2001; 16(2) 406-416

24. Frost ML et al. Contact quantitative ultrasound: an evaluation of precision, fracture discrimination, age-related bone loss, and applicability of the WHO criteria. Osteoporosis International 1999; 10: 441-449

25. Ayers M et al. Reconciling quantitative ultrasound measurements of the calcaneus with X-ray based measurements of the central skeleton. Journal of Bone and Mineral Research 2000; 15(9) 1850-1855

26. Stewart A, Reid DM. Quantitative ultrasound or clinical risk factors – which best identifies women at risk of osteoporosis? British Journal of Radiology 2000; 73: 165-171

27. Pocock et al. Screening for osteoporosis: what is the role of heel ultrasound? MJA 1996; 164: 367-370

28. Miller P et al. Prediction of Fracture Risk in Postmenopausal Caucasian Women With Peripheral Bone Densitometry: Evidence From the NORA. Journal of Bone and Mineral Research 2001; 16(Supp 1) S193 [abstract].

29. North American Menopause Society: *A decision tree for the use of estrogen replacement therapy or hormone replacement therapy in postmenopausal women: consensus opinion of The North American Menopause Society.* Menopause 2000; 7: 76-86

30. North American Menopause Society: *Management of postmenopausal osteoporosis: position statement of The North American Menopause Society.* Menopause 2002; 9: 84-101

31. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristics (ROC) curve. Radiology 1982; 143: 29-36

32. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. Radiology 1983; 148: 839-843

33. Delong ER et al. Comparing the areas under two or more correlated receiver operating characteristics curves: a non-parametric approach. Biometrics 1988; 44: 837-845

34. Zeger S et al. Models for longitudinal data: A likelihood approach. Biometrics 1988; 44: 1049-60.

35. Zeger S, Liang L. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13-22.

36. Henry YM, Eastell R. Ethnic and gender differences in bone mineral density and bone turnover in young adults: effect of bone size. Osteoporos International 2000;11(6):512-517

37. Bohannon AD. Osteoporosis and African American women. J Womens Health Gend Based Med 1999 Jun;8(5):609-615