Use of Clinical Prediction Rules in Detecting Osteoporosis in a Population-Based Sample of Postmenopausal Women

Karen F. Mauck, MD, MSc; Maria-Teresa Cuddihy, MD, MPH; Elizabeth J. Atkinson, MSc; L. Joseph Melton III, MD, MPH

**Background:** Osteoporosis clinical prediction rules attempt to identify the postmenopausal women in whom, on the basis of individual risk factors, bone densitometry will detect low bone mass. We assessed and compared the diagnostic properties of the following 3 osteoporosis clinical prediction rules: the Simple Calculated Osteoporosis Risk Estimation, Osteoporosis Risk Assessment Instrument, and National Osteoporosis Foundation practice guidelines.

**Methods:** Secondary data analysis of an existing population-based sample of postmenopausal women 45 years or older (N = 202) in Rochester, Minn.

**Results:** Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values, and positive (LR+) and negative (LR−) likelihood ratios were calculated using the World Health Organization diagnosis of osteoporosis as the reference standard. The Simple Calculated Osteoporosis Risk Estimation had a sensitivity of 100%, specificity of 10%, PPV of 27%, NPV of 100%, LR+ of 1.4, and LR− of 0. The Osteoporosis Risk Assessment Instrument had a sensitivity of 98%, specificity of 40%, PPV of 29%, NPV of 77%, LR+ of 1.4, and LR− of 0.4. The National Osteoporosis Foundation practice guidelines had a sensitivity of 100%, specificity of 10%, PPV of 27%, NPV of 100%, LR+ of 1.1, and LR− of 0. The Simple Calculated Osteoporosis Risk Estimation and Osteoporosis Risk Assessment Instrument were much more specific in postmenopausal women younger than 65 years compared with those 65 years or older.

**Conclusions:** Our results suggest that these clinical prediction rules do not perform well as a general screening method to identify postmenopausal women who are more likely to have osteoporosis; however, the Osteoporosis Risk Assessment Instrument and Simple Calculated Osteoporosis Risk Estimation may be useful in identifying some women who need not undergo testing, especially younger postmenopausal women.

Arch Intern Med. 2005;165:530-536
disagreement about how this individualized approach to screening should be achieved. Most guidelines recommend using risk factors to select patients for BMD testing, but because of inadequate data, there is no consensus on which risk factors to use. Although hundreds of studies report clinical risk factors for osteoporosis and fractures, most of these are epidemiological studies. Very few studies evaluate how to use these risk factors to identify individual women who are at risk for fracture. Clinicians are frequently faced with decisions about BMD testing in individual postmenopausal patients, most of whom have at least 1 of the several risk factors associated with increased risk of fracture. From the need to make this individualized approach to osteoporosis screening more practical and simplified for clinicians, several clinical prediction rules (CPRs) have been developed.

Clinical prediction rules are designed to assist medical decision making. They stratify patients into risk subgroups on the basis of differing probabilities of disease as determined by summarizing risk factors with a point system. A good CPR will differentiate women who are more likely to have low BMD from those who are more likely to have normal BMD when they undergo testing. Many CPRs have been developed, several that attempt to predict BMD outcomes and several that attempt to predict fracture outcomes. The quality of most of these CPR development studies has been rated fair or poor, most often because of methodological limitations. A system of the Rochester Epidemiology Project has been applied to different populations with inconsistent results, and the ORAI has never been validated in a population independent of that in which it was developed.

METHODS

STUDY DESIGN AND SAMPLE POPULATION

We performed a secondary data analysis of an existing population-based cohort of postmenopausal women in Rochester who are participating in an ongoing, prospective study designed to assess osteoporosis prevalence, risk factors, and outcomes. These women were recruited from an age-stratified random sample of Rochester women using the medical records linkage system of the Rochester Epidemiology Project. More than half of the Rochester population is seen annually in this system, and most are attended in any 3-year period. Thus, the enumerated population (Rochester women seen in 1990 ± 1 year) approximates the underlying population of the community, including free-living and institutionalized individuals. A total of 938 women 20 years and older were approached for this study, but 126 were ineligible (89 were deemed and could not give informed consent; 11 were pregnant; 9 were radiation workers; 8 were participants in an ongoing clinical trial of osteoporosis prophylaxis; and 9 died before they could be contacted). Of the 812 eligible women, 351 (43%) participated. About 30 women per decade of age from 20-29 years to older than 80 years were enrolled. Participation rates per age group were 50%, 48%, 56%, 65%, 57%, 39%, and 22%, respectively. All women provided written informed consent before entry into the study, which was conducted after approval from the Mayo Foundation Institutional Review Board. The subjects included 138 premenopausal women (mean ± SD age, 35.0 ± 8.6 years, range, 21-54 years) and 213 postmenopausal women (mean ± SD age, 67.8 ± 13.2 years, range, 34-93 years). For the present analysis, we used the baseline data and BMD measurements from the postmenopausal subset of women 45 years and older (N = 202).

Baseline data were collected by a single study nurse using a structured interview that was complemented by review of each subject’s complete (inpatient and outpatient) medical record in the community. The study nurse also measured each subject’s height with a stadiometer and weight (light clothing without shoes) on a balance-beam scale. All subjects then underwent BMD testing performed by a single research technician. Areal BMD was determined for the femoral neck and the anterior-posterior lumbar spine using dual-energy x-ray absorptiometry (QDR2000 instrument; Hologic, Waltham, Mass) and QDR2000 software version 5.40. The coefficient of variation was 1.5% for the femoral neck BMD measurement and 0.6% for the anterior-posterior spine.

APPLICATION OF THE CPRs

The description of each CPR and its respective scoring system is outlined in Table 1. Each CPR has a prespecified number of points assigned by the developing authors as the cutoff point to be used for recommending BMD testing. For the SCORE, a result of 6 points or more was needed to recommend a woman for BMD testing, whereas a total of 9 or more points was needed.
Table 2. Summary of Descriptive Characteristics, Osteoporosis Risk Factors, and BMD of an Age-Stratified Sample of 202 Postmenopausal Women in Rochester, Minn

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>No. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Osteoporosis risk factors</strong></td>
<td></td>
</tr>
<tr>
<td>Estrogen use</td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>120 (59)</td>
</tr>
<tr>
<td>Ever</td>
<td>82 (41)</td>
</tr>
<tr>
<td>Current</td>
<td>43 (21)</td>
</tr>
<tr>
<td>Fracture history*</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>42 (21)</td>
</tr>
<tr>
<td>No</td>
<td>160 (79)</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>118 (58)</td>
</tr>
<tr>
<td>Ever</td>
<td>84 (42)</td>
</tr>
<tr>
<td>Family history of fracture†</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>13 (6)</td>
</tr>
<tr>
<td>No</td>
<td>189 (94)</td>
</tr>
<tr>
<td><strong>BMD</strong></td>
<td></td>
</tr>
<tr>
<td>Femoral neck T score‡</td>
<td></td>
</tr>
<tr>
<td>Greater than −2.0</td>
<td>95 (47)</td>
</tr>
<tr>
<td>−2.0 or less</td>
<td>107 (53)</td>
</tr>
<tr>
<td>−2.5 or less</td>
<td>69 (34)</td>
</tr>
<tr>
<td>Age 45-64 y</td>
<td>11 (5)</td>
</tr>
<tr>
<td>Age ≥65 y</td>
<td>58 (29)</td>
</tr>
<tr>
<td>Lumbar spine T score‡</td>
<td></td>
</tr>
<tr>
<td>−2.5 or less</td>
<td>15 (7)</td>
</tr>
<tr>
<td>Age 45-64 y</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Age ≥65 y</td>
<td>12 (6)</td>
</tr>
</tbody>
</table>

Abbreviation: BMD, bone mineral density.
*Includes history of at least minimal trauma fracture of the wrist or forearm, hip, or spine after age 45 years.
†Includes history of osteoporotic fracture in first-degree relative.
‡Calculated using BMD mean and standard deviation for Rochester women (aged 20-29 years) as the young healthy reference population. Categories are not mutually exclusive.

The study sample included 202 postmenopausal women (99% white) with mean age of 69.2 ± 11.9 years and mean weight of 67.3 ± 13.4 kg. The mean BMD at the femoral neck was 0.67 ± 0.12 g/cm², and 34% of the women had osteoporosis as defined by the WHO reference standard (femoral neck BMD T score, −2.5 or less). The age-adjusted (to 1990 US white women aged ≥45 years) prevalence was 25%. If the lumbar spine measurements were used as the reference standard to establish the diagnosis of osteoporosis, only 7% of the women in this sample would have osteoporosis. Thus, we report additional analysis only for the femoral neck. A summary of the descriptive characteristics, osteoporosis risk factors that were used in the CPRs, and BMD measurements of the sample is provided in Table 2.

In the unadjusted analysis, all 3 rules were found to be quite sensitive, but the specificities were not optimal for the ORAL.\textsuperscript{11} The NOF recommendations were not intended to be used as a scoring system, but for the purposes of comparison we used the point system devised by Cadarette et al\textsuperscript{16} in which each NOF factor was assigned 1 point. Anyone with 1 point or more was recommended for BMD testing.

Operating characteristics, including sensitivity, specificity, positive (PPV) and negative (NPV) predictive values, and positive (LR+) and negative (LR−) likelihood ratios, were calculated for each CPR using the recommended cutoff points. The reference standard for a positive test result was established using the World Health Organization definition of osteoporosis, specifically, femoral neck BMD level 2.5 SDs below the mean for healthy young women.\textsuperscript{39} The reference mean femoral neck BMD (0.903 g/cm²; SD, 0.119 g/cm²) for healthy young women was determined using Rochester women aged 20 to 29 years; the reference mean anterior-posterior lumbar spine BMD was 1.084 g/cm² (SD, 0.135 g/cm²).\textsuperscript{39}

**STATISTICAL ANALYSIS**

Basic demographic data and subject characteristics were reported as mean ± SD for continuous variables and as percent-

ages for categorical variables. Sensitivity, specificity, and their corresponding 95% confidence intervals (CIs) were calculated at the recommended cutoff point for each CPR. Sensitivity was defined as the proportion of women with osteoporosis at the femoral neck (T score of −2.5 SD or less) for whom the decision rule recommended BMD testing. Specificity was defined as the proportion of women without osteoporosis at the femoral neck (T score of greater than −2.5 SDs) for whom the decision rule did not recommend testing. Because sensitivity and specificity are not that useful to the clinician, who is more interested in the likelihood that a patient has osteoporosis after having used the decision rule, we also calculated the PPV, NPV, LR+, and LR− for each CPR to assess its clinical usefulness. The PPV was defined as the proportion of subjects with osteoporosis who had a positive test result or true-positive findings (positive test result and osteoporosis) divided by the number of subjects with a positive test result. The NPV was defined as the proportion of subjects without osteoporosis who had a negative test result or true-negative findings (negative test result and no osteoporosis) divided by the number of subjects with a negative test result. The LR+ was defined as sensitivity/(1-specificity), that is, the probability of a positive test result in subjects with osteoporosis divided by the probability of a positive test result in subjects without osteoporosis. The LR− was defined as (1-sensitivity)/specificity, that is, the probability of a negative test result in subjects with osteoporosis divided by the probability of a negative test result in subjects without osteoporosis. We calculated 95% CIs using the method described by Simel et al.\textsuperscript{40}

Receiver operating characteristic (ROC) curves were constructed for each CPR using the method of DeLong et al,\textsuperscript{41} and the areas under the ROC curves were statistically compared. By using the method of Obuchowski\textsuperscript{42} for comparing these CPRs in the same set of 202 subjects (with a correlation between the areas estimated to be 0.7), we had greater than 90% power to detect a difference of 0.1 between these tools with respect to diagnostic accuracy. In addition, we also constructed age-adjusted ROC curves using a bootstrap method to compare the areas under the ROC curves.

Because our sample was stratified by age, it is not, per se, representative of the general population. Therefore, the data were directly age adjusted to the population structure of US white women 45 years and older in 1990. After this adjustment, we recalculated all operating characteristics and CIs.

©2005 American Medical Association. All rights reserved.
The SCORE had an overall sensitivity and specificity of 100% (95% CI, 95%-100%) and 25% (95% CI, 18%-33%), respectively, with a PPV of 41% (95% CI, 34%-49%) and NPV of 100% (95% CI, 95%-100%). The LR+ and LR− were 1.1 (95% CI, 1.0-1.2) and 0.0 (95% CI, not applicable), respectively. The ORAI had an overall sensitivity and specificity of 89% (95% CI, 89%-100%) and 36% (95% CI, 28%-44%), respectively, with a PPV of 36% (95% CI, 24%-51%) and NPV of 100% (95% CI, 95%-100%). The LR+ and LR− were 1.3 (95% CI, 1.2-1.5) and 0.0 (95% CI, not applicable), respectively. The NOF had an overall sensitivity and specificity of 99% (95% CI, 94%-100%) and 89% (95% CI, 83%-95%), respectively, with a PPV of 99% (95% CI, 98%-100%) and NPV of 100% (95% CI, 95%-100%). The LR+ and LR− were 1.5 (95% CI, 1.4-1.7) and 0.0 (95% CI, not applicable), respectively.

We also report in Table 3 the operating characteristics of each tool stratified by age group (45-64 vs ≥65 years). For the SCORE and ORAI, the specificity was higher in the younger postmenopausal women compared with the older women. All of the CPRs were quite sensitive in both subgroups. The PPVs for all 3 CPRs were higher in the younger subgroup compared with the older subgroup, whereas the NPVs of all 3 CPRs were similar in the younger and older subgroups.

Because our age-stratified sample had a higher percentage of elderly women than the general population, we directly age adjusted the data to the 1990 US population of white women 45 years and older. Given the resulting increase in the influence of younger women, these CPRs were a bit more specific in identifying women with osteoporosis, whereas the sensitivities changed only minimally. The PPVs of all 3 CPRs were poor in the age-adjusted sample, whereas the NPVs remained high (Table 3).

The 3 CPRs were further compared using areas under the ROC curves (Table 4). In the unadjusted analysis, the SCORE had an area under the ROC curve of 0.87 (95% CI, 0.81-0.92); the ORAI had an area of 0.84 (95% CI, 0.78-0.89); and the NOF guideline had an area of 0.70 (95% CI, 0.63-0.77). There was no difference between the areas under the ROC curves for the SCORE and the ORAI (P = .13). The SCORE and the ORAI were found to have better discriminatory performance than the NOF guideline (P < .001 for both comparisons). To further explain.
plore the accuracy of these tools in younger vs older women, we also calculated ROC curves for each CPR for the subset of women aged 45 to 64 years compared with those 65 years or older. For women aged 45 to 64 years (n=79), there was no difference between the areas under the ROC curves for the 3 CPRs (P=.23). For the subset of women 65 years or older (n=123), there was no difference between the areas under the ROC curve for the SCORE and the ORAI (P=.77), but the SCORE and ORAI had better discriminatory performance than the NOF guidelines (P<.001 and P=.009, respectively).

We also compared the areas under the ROC curves that were derived using an age-adjusted analysis (Table 4). The SCORE had an area of 0.85 (95% CI, 0.80-0.89); the ORAI had an area of 0.79 (95% CI, 0.74-0.83); and the NOF guidelines had an area of 0.65 (95% CI, 0.58-0.71). There was a statistically significant difference between the areas under the curve for the SCORE compared with the ORAI and between the SCORE and ORAI compared with the NOF guidelines.

We applied the SCORE, ORAI, and NOF guidelines to an independent, population-based sample of postmenopausal women to compare the operating characteristics and to determine the clinical usefulness of each tool as an osteoporosis screening method in the general population. Our results suggest that these CPRs do not perform well as a general screening method to identify postmenopausal women who are more likely to have osteoporosis; however, the ORAI and SCORE may be useful in identifying some women who need not undergo testing, especially among younger postmenopausal women.

A good screening test must have a high sensitivity, so that it does not miss the few cases of disease that are present, and a high specificity, to reduce the number of people with false-positive results who require further diagnostic testing. The CPRs were quite sensitive in our population-based sample, but none was very specific. The SCORE would recommend BMD testing in 71% of women without osteoporosis, whereas the ORAI would recommend testing in 60% of women without osteoporosis. The NOF guidelines were the least specific in that they would recommend BMD testing in 90% of women without osteoporosis. However, in the younger postmenopausal women, the SCORE and ORAI recommended BMD testing in only 50% and 31% of women without osteoporosis, respectively, while maintaining a favorably high sensitivity.

The findings we report herein are similar to those reported by Cadarette et al,38 who compared the SCORE, ORAI, and NOF practice guidelines in a large sample of Canadian women and concluded that the SCORE and ORAI were superior to the NOF guidelines in correctly identifying high-risk women for BMD testing. Although that study included a large sample of postmenopausal women, it was limited in that the sample was not likely representative of the general population. Women who were taking hormone therapy for less than 5 years were excluded, and older age groups were oversampled without statistical adjustment in the analysis. Those authors reported sensitivities very similar to ours, but the specificities were lower in their study, likely reflecting an underrepresentation of younger women in their sample.

Our results are also useful to help clarify some of the inconsistent results that have been shown with the SCORE. Since its development in 1998, the SCORE has been evaluated in several studies, with inconsistent results,20,32-35 likely owing to differences in the sample populations. A CPR should be validated in a population similar to that in which it is meant to be used—in the case of a CPR that will be used as an osteoporosis screening tool in the general population of postmenopausal women, the CPR should ideally be tested in a sample that is representative of this population. Lydick et al32 described a sensitivity and specificity of 89% and 50%, respectively, in the development cohort (mean age, 61.5 years) and 91% and 40%, respectively, in the validation cohort (mean age, 63.1 years), using a T score of −2.0 or less at the femoral neck as the definition of osteoporosis. The development and validation cohorts had osteoporosis prevalences of 38% and 44%, respectively. However, when the SCORE was later applied to a sample of community-dwelling postmenopausal women in Rancho Bernardo, Calif (mean age, 72.5 years), with osteoporosis prevalence of 67% as defined by BMD T score of −2.0 or less at the femoral neck,32 the sensitivity was 98% and the specificity was only 12.5%. Cadarette et al38 applied the SCORE to a sample of 398 postmenopausal women participating in the large population-based Canadian Multicenter Osteoporosis Study. Those authors reported a sensitivity of 90% and a specificity of 32% in their population of postmenopausal women (mean age, 64.5 years) with an osteoporosis prevalence of 50% (using BMD T score of −2.0 or less at the femoral neck as the definition of osteoporosis). More recently, Cadarette et al39 applied the SCORE to an even larger (N=2365) sample of the same study population (mean age, 66.4 years) and reported a sensitivity of 99.6% and a specificity of 17.9% (using BMD T score of −2.0 or less as the definition of osteoporosis, this sample had a prevalence of 25.4%; using a BMD T score of −2.5 or less, the prevalence was only 10%). Two other studies have also reported sensitivities and specificities for the SCORE; however, these sample populations consisted of postmenopausal women who were referred for BMD testing and thus were not representative of a true situation for which this tool would be used.34,35 These inconsistencies can best be explained by the spectrum of patients to whom the SCORE was applied. The Rancho Bernardo study was a sample of community-dwelling women; however, there must have been patient selection bias because the prevalence of osteoporosis was so high and the average age was 10 years older than those in the original study by Lydick et al.20 As patient age increases, the more likely the SCORE will produce false-positive results, because age is a major point contributor in the SCORE calculation, thus explaining why the specificity reported from the Rancho Bernardo study was so much lower than that reported by Lydick et al.20 The 2 studies by Cadarette et al33,38 nicely illustrate this point as well. The first study33 with 398 women had an average age similar to the co-
horts studied by Lydick et al. However, in the larger study of 2365 women, the mean age was older and the specificity was lower. In our study, because the cohort was stratified by age, there was overrepresentation of the older women compared with that of the normal population, and therefore the mean age was 69.2 years. When the data were directly age adjusted, the sensitivity and specificity were 100% and 27%, respectively. It is difficult, however, to compare our results directly with those of the other studies because the BMD T score cutoff used to define osteoporosis is −2.0 or less in most of those studies, and we used the more conservative value of T score of −2.5 or less. In our study sample, the age-adjusted prevalence of osteoporosis as defined by BMD T score of −2.5 or less or at the femoral neck was 25%, compared to a similar 22% figure for white women from the Third National Health and Nutrition Examination Survey data. The Third National Health and Nutrition Examination Survey reports only prevalence of osteoporosis (BMD T score of −2.5 or less) and osteopenia (BMD T score of −1 or less to −2.5); it does not report a prevalence for a BMD T score of −2.0 or less to compare with those of the other SCORE studies.

Our data clarify how these CPRs may best be applied in a clinical setting. In light of the recent US Preventive Service Task Force recommendation that all women 65 years and older should undergo screening for osteoporosis, and the fact that Medicare now reimburses for BMD testing in estrogen-deficient women with osteoporosis risk factors, there is less ambiguity for the clinician about osteoporosis screening in this age group. However, for those postmenopausal women who are younger than 65 years, there remains uncertainty as to who should undergo screening. The SCORE and the ORAI perform better in the group of women aged 45 to 64 years and may help clinicians to identify those younger postmenopausal women who are higher risk. However, on the basis of the NPVs and LR− in Table 3, it is evident that these tools are more helpful in that they identify those few women who do not need BMD testing rather than those who do.

Although our study was limited by a small sample size, we believe this sample closely represents our population of postmenopausal women in whom these osteoporosis CPRs would be applied. However, one must keep in mind that even if these CPRs are useful in the general population, translating the results of population studies to the care and management of individual patients requires additional consideration. The generalizability of these tools is currently limited until their applicability in the clinical setting can be tested in a prospective manner. In addition, most of these tools have been developed and tested in rather homogeneous cohorts of non-Hispanic white women, which obviously limits their use in more heterogeneous populations.

CONCLUSIONS

The limited performance of the currently available osteoporosis CPRs raises question about the clinical usefulness of such tools as an effective method for screening for osteoporosis. However, they may prove useful in aiding clinicians to select those few women who are younger than 65 years and have a negative CPR outcome, who do not need to be referred for BMD testing. Further studies in various clinical settings with more heterogeneous populations are warranted to prospectively validate the SCORE and the ORAI.

Accepted for Publication: August 25, 2004.

Correspondence: Karen F. Mauck, MD, MSc, Mayo Clinic, 200 First St SW, Rochester, MN 55905 (mauck.karen@mayo.edu).

Funding/Support: This study was supported in part by research grant AR 27065 from the National Institutes of Health, US Public Health Service, Bethesda, Md.

Previous Presentation: Portions of this study were presented in poster form at the Fifth International Symposium on Clinical Advances in Osteoporosis; March 6-9, 2002; Honolulu, Hawaii.

Acknowledgment: We thank Mark Liebow, MD, for his editorial support and valuable input regarding clinical prediction rules.

REFERENCES

16. Elliot JR, Gilchrist NL, Wells JE, Aylind E, Turner J, Sainsbury R. Historical as


