BoneKEy. 2007 August;4(8):226-230
http://www.bonekey-ibms.org/cgi/content/full/ibmske;4/8/226
DOI: 10.1138/20070269

COMMENTARIES

Predicting Fracture Risk: Tougher Than It Looks

Warren S. Browner

California Pacific Medical Center Research Institute, San Francisco, California, USA

Commentary on: Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, Burckhardt P, Cooper C, Christiansen C, Cummings S, Eisman JA, Fujiwara S, Glüer C, Goltzman D, Hans D, Krieg MA, La Croix A, McCloskey E, Mellstrom D, Melton LJ, Pols H, Reeve J, Sanders K, Schott AM, Silman A, Torgerson D, van Staa T, Watts NB, Yoshimura N. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int*. 2007 Aug;18(8):1033-46.

A few months ago, *Osteoporosis International* (*OI*) published an interesting paper by Kanis *et al*. on the prediction of hip and other osteoporotic fractures in people aged 50 years and older (1). They used information from nine large epidemiologic studies to develop their risk models and then validated their results in 11 other studies to determine the effect of adding a few clinical risk factors, such as body mass index and use of glucocorticoids, to models based on bone mineral density (BMD) alone. What were their main findings? First, models for predicting hip fractures were substantially better than those for other osteoporotic fractures, whether the models included BMD alone, clinical risk factors only, or both. Second, BMD was a remarkably strong predictor of hip fractures, especially for those under 70 to 75 years of age. Third, both BMD and clinical risk factors were better at predicting hip fractures among younger people (ages 50 to 60) than among those who were older (ages 80 to 90). Fourth, for hip fracture, clinical information added significantly (in the statistical sense) to models based on BMD alone. But the addition of clinical risk factors did not really improve risk prediction all that much, particularly among those ages 70 and older, in whom the vast majority of hip fractures occur. On the other hand, for predicting other (non-hip) osteoporotic fractures, clinical risk factors were more useful than BMD, but even the best models for these fractures were not very impressive. As it turns out, fracture risk prediction, for a number of reasons to be discussed here, is much more difficult than it first appears.

Risk Models in General

Were these results—particularly the modest additional value of clinical risk factors for predicting hip fracture among those older than age 70, and the models' modest predictive ability for non-hip fractures—disappointing or, well, predictable? To answer that question, we will begin with the general problem of risk stratification (or prediction) models, and then return to the specific example of predicting fracture from clinical information and bone density.

Risk prediction models have as a goal distinguishing those at high risk from those at low risk. However, many of the analytic tools used to identify the variables that go into risk models were developed to suggest possible risk factors (*i.e.*, causes) for an outcome. The criteria used to select a risk factor—rejecting the null hypothesis of no association between that variable and the outcome, and then determining whether that association is confounded by other variables—don't necessarily identify characteristics that are useful for risk prediction.

Developing risk prediction models is much more difficult than showing that a risk factor, or set of risk factors, has a statistically significant association with an outcome. Indeed, even a highly significant association—one with a vanishingly small P value—may have almost no discriminatory ability. Somewhat ironically, given the emphasis on big studies, P values are least useful when the sample size is large—because even a very small difference between two groups can result in a significant P value. (Being able to predict a coin toss 51% of the time can make you rich in the long run, but won't impress someone who watches your performance for "only" a few hundred flips).

So what should a P value be used for? The P value (or seeing whether the confidence interval overlaps 1.0) has a simple job—to identify a statistically significant difference. Without that, a risk prediction model has little meaning. But that's all a P value does: it's the "ante" required to play the hand.

Second, most outcomes, like hip fractures, are rare, particularly during relatively short follow-up times. Thus, in many situations, our ability to predict who is going to have a particular outcome is not all that much better than a rule that goes "just say no one is going to have the outcome." With an annual hip fracture risk of 0.1%, as in the first Gothenburg cohort (2), that latter rule would be right 999 times per 1000 guesses.

Next, while there may be several (or even many) statistically significant risk factors for an outcome, not very many of them are "strong" risk factors with large relative risks (RRs). Indeed, having a low BMD is one of the strongest risk factors in routine clinical use. Moreover, even if you could miraculously classify everyone into one of just two groups—one at low risk (say, a 1% risk over ten years) and the other at high risk (with a 10% risk during that time)—that isn't really all that helpful. After all, nine in ten members of the high-risk group will not have the outcome and you still have not identified the one in ten who will. And of course, even if you can develop a model that generates a ten-fold risk gradient, not everyone will be either high- or low-risk: most will fall somewhere in the middle. Thus what matters is the performance of a risk score through its entire range.

To evaluate that entire range, we use another statistic—the area under a receiver-operator characteristic (ROC) curve. Developed in engineering, and often used for studies of diagnostic tests, ROC curves (like those in Figure 2 of the *OI* article) plot the sensitivity of a test at various cut points for being called "positive" versus one minus the test's specificity at those cut points. A reminder: sensitivity is defined as "positivity in disease" (PID)—in this case, how often the risk score was greater than the cut point among those with a fracture. Specificity is defined as "negativity in health" (NIH)—in this situation, how often the risk score was less than that cut point among those without a fracture. For those who cannot remember these definitions, the mnemonics are PID (patients are *sensitive* about having Pelvic Inflammatory Disease) and NIH (the National Institutes of Health are *specific* about whom they fund—recently, almost no one!). The ROC curve for a perfect test or risk score approaches the upper left (northwest) corner of the graph; the curve for a useless test lies close to the southwest to northeast diagonal. Since the graph represents a "one unit by one unit square," the area under an ROC curve (*i.e.*, the area between the curve and the x-axis) approaches 1.0 as the curve nears the northwest corner. On the other hand, the area approaches 0.5 as the curve nears the diagonal and the risk score becomes no better than a random guess. Risk scores that have a "substantial" area under the curve (as in the leftmost graph in Figure 2 of the *OI* article) do much better than chance.

Back to Bone

Now let's return to the example of predicting fracture risk from BMD and clinical information such as family history and use of glucocorticoids. Kanis *et al*. developed six basic models: three for predicting hip fracture (one model with BMD alone, one with clinical risk factors alone, and one with both BMD and clinical risk factors), and the same for non-hip fractures. They used these models in five different age groups (from 50

BoneKEy. 2007 August;4(8):226-230
http://www.bonekey-ibms.org/cgi/content/full/ibmske;4/8/226
DOI: 10.1138/20070269

years to 90 and older, by 10-year groups), for a total of 30 different models. For all of these situations, they assigned each participant a risk score (on a standardized Z scale) and determined the association between the score and two outcomes—hip fracture and other osteoporotic fractures—to calculate the relative risks per SD increase in the risk score, along with its confidence intervals. All of these relative risks were greater than 1.0 and highly statistically significant (*i.e.*, none of the confidence intervals came close to overlapping 1.0). Thus the ante to play was covered.

So how well did the models do? Assuming that the risk scores are normally distributed (as appears to be the case; see Figure 1 in the *OI* paper), there is a one-to-one correspondence between the relative risk per SD change and the area under the ROC curve (Figure 1 in this *Commentary*), as recognized by the authors of the *OI* article. Only large relative risks—I'd say at least 3 per SD, corresponding to a "substantial" ROC area of about 78%—have valuable discriminatory ability. BMD alone, at least for hip fractures among those under 60 or 65 years of age, is pretty good in this regard (see Table 3 of the *OI* article).

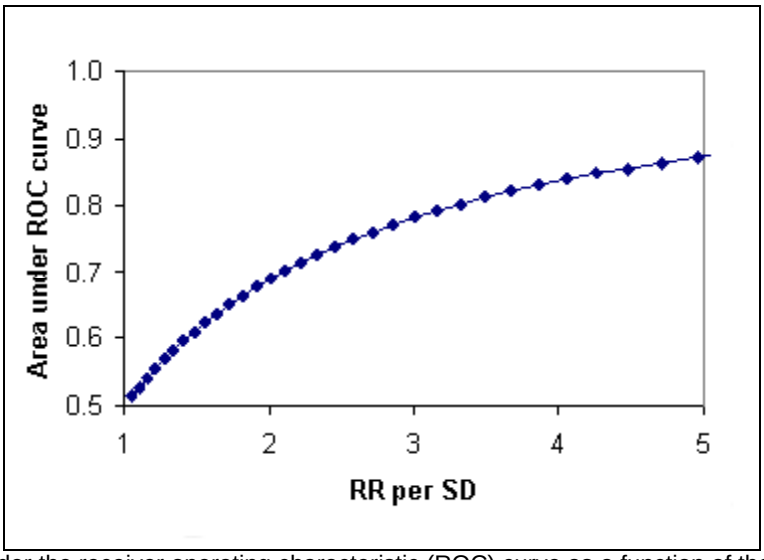

Figure 1. Area under the receiver operating characteristic (ROC) curve as a function of the relative risk (RR) per SD increase in risk score. See Appendix to reference 1 for details.

Relative risks and areas under ROC curves have another problem—they indicate the "operating characteristics" (or risk gradient) of the risk score itself, but they ignore the absolute risk of what is being predicted. The clinical usefulness of a risk prediction model, however, does depend on that risk. The utility of a model is greater as the proportion of those at high risk who actually develop that outcome increases. (The opposite should be true for those classified as being at low-risk, of course).

So what should we do after being let down by P values, relative risks, and even ROC areas? Some have suggested using yet another type of statistic to indicate how often a new prediction rule successfully reclassifies people from one risk class to another (3). This has its own problems—such as how to define "successful reclassification" and "risk class"—but it does provide an easy-to-interpret statistic, namely the percentage of subjects who are reclassified successfully. But this approach also has it limitations.

Let's review two different ways of seeing whether a risk classification system improves risk prediction. The first method asks "What percentage of all events occurs among those in the top decile," comparing

BoneKEy. 2007 August;4(8):226-230
http://www.bonekey-ibms.org/cgi/content/full/ibmske;4/8/226
DOI: 10.1138/20070269

the new with the old scoring system. The second method asks "What percentage of patients is classified as being at 'high risk,' defined as at least twice as great as the overall risk," again comparing the old and the new scores.

In this regard, the *OI* models are incomplete—the authors did not provide the absolute risk (say, per 1000 person-years) of fracture, just the relative risk. Nor did they provide the "inner workings" of the model or determine how well their models did at identifying high- (or low-) risk people, by comparing the observed rates with the expected rates in those groups. So we can't really answer the questions, but some guesstimates are provided in Figure 2 of this *Commentary*. These estimates are based on the *OI* results, assuming an underlying hip fracture risk in the population of 10%, roughly the 10-year risk in ambulatory white women ages 65 years and older in the U.S. (4). For example, suppose the risk score based on BMD alone has an RR per SD of 3.68 and the score from the model based on combining BMD and clinical risk factors has an RR per SD of 4.23 (see Table 3 of the *OI* paper, for age 50 years). Then the model that combines BMD with clinical information increases the percentage of all hip fractures that occur in the top decile of risk scores from about 33.6% to about 35.0%. Alternately, the combined model raises the percentage of high-risk patients (those classified as having at least twice the underlying risk) from about 14.5% to about 15.5%. Thus, even a fairly substantial increase in the relative risk leads to a small increase in the ability to identify those at high risk. At older ages (*e.g.*, age 70), the increase in the relative risk from adding clinical risk factors to the *OI* model is quite modest (2.78 to 2.91 per SD), and very few patients are reclassified. Can we all just say "Argh"?

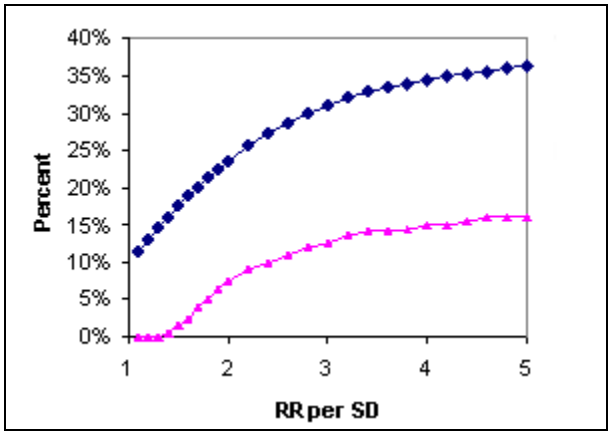

Figure 2. Percent of all outcomes that occur among persons in the highest decile of risk (blue diamonds) and the percent of all persons whose risk is at least twice the overall risk (pink triangles) as functions of the relative risk (RR) per SD increase in risk score, assuming an overall risk of 0.1.

Why Have Risk Prediction Models Anyway?

The main purpose of developing these sorts of models is to identify patients at high enough risk to warrant an intervention that is either too costly or too risky (or both) to apply to everyone. (In other situations, like survival after admission to an intensive care unit (ICU), prediction models are useful for determining whether a particular ICU has a higher- or lower-than-expected mortality). Success in classifying someone as being high-risk depends entirely on what happens as a result of that classification. Presumably, those at high risk would be encouraged to start preventive treatment, and presumably the treatment is (at least as) efficacious in those at high risk as it is in everyone else. To the extent that a treatment is less efficacious in, or less acceptable to, high-risk patients, it becomes less important to

identify them (5). Thus, the real "gold standard" in evaluating risk prediction models would require showing that those who are risk-stratified (some of whom are subsequently treated) have fewer adverse outcomes, lower costs, or perhaps both, than those who received usual care. This may explain why we have abandoned the gold standard and use paper currency.

Conflict of Interest: The author reports that no conflict of interest exists.

References

1. Kanis JA, Oden A, Johnell O, Johansson H, De Laet C, Brown J, Burckhardt P, Cooper C, Christiansen C, Cummings S, Eisman JA, Fujiwara S, Glüer C, Goltzman D, Hans D, Krieg MA, La Croix A, McCloskey E, Mellstrom D, Melton LJ, Pols H, Reeve J, Sanders K, Schott AM, Silman A, Torgerson D, van Staa T, Watts NB, Yoshimura N. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. *Osteoporos Int*. 2007 Aug;18(8):1033-46.

2. Svanborg A. Seventy-year-old people in Gothenburg a population study in an industrialized Swedish city. II. General presentation of social and medical conditions. *Acta Med Scand Suppl*. 1977;611:5-37.

3. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006 Jul 4;145(1):21-9.

4. Cummings SR, Nevitt MC, Browner WS, Stone K, Fox KM, Ensrud KE, Cauley J, Black D, Vogt TM. Risk factors for hip fracture in white women. Study of Osteoporotic Fractures Research Group. *N Engl J Med*. 1995 Mar 23;332(12):767-73.

5. Browner WS. Estimating the impact of risk factor modification programs. *Am J Epidemiol*. 1986 Jan;123(1):143-53.