Published in final edited form as:
J Natl Cancer Inst. 2008 July 16; 100(14): 978–979. doi:10.1093/jnci/djn215.

Gauging the Performance of SNPs, Biomarkers, and Clinical Factors for Predicting Risk of Breast Cancer

Margaret S. Pepe and **Holly E. Janes**
Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA (MSP, HEJ).

Predicting risk of cancer for individuals has long been a goal of medical research. If an individual's risk could be predicted, then prevention and screening modalities could be targeted toward those at meaningfully high risk. This approach is not only more cost efficient than targeting the whole population but also more ethical, at least when interventions are burdensome to the individual. The quest for risk predictors has been revitalized with the emergence of technologies that measure genetic information and other molecular and physiological attributes of the individual. In this issue of the Journal, Gail (1) asks to what extent newly discovered associations between seven single-nucleotide polymorphisms (SNPs) and incidence of breast cancer can improve assessment of breast cancer risk. Comparisons are made with models that employ standard clinical factors to evaluate the incremental value of the SNPs for prediction over the standard clinical information. Using estimated relative risks and allele frequencies, Gail finds that the SNPs are expected to have a small effect on the capacity of prediction models to distinguish women who will and will not develop breast cancer. Because he assumes best-case scenarios, his results probably provide upper limits on expected increments in risk prediction with SNPs. He postulates that many more SNPs with these levels of association with breast cancer will need to be discovered to substantially improve risk prediction. Gail's arguments demonstrate that the sample sizes needed to discover an adequate number of SNPs will need to be very large indeed. Although his calculations are based on many assumptions, they provide a good place from which to start the discussion about what types of markers and studies will be needed to make progress in this field.

Gail uses the area under the receiver operating characteristic (ROC) curve (AUC) to summarize and compare prediction models. Although this is a popular statistical approach with a long history (2), there has been considerable criticism of it, with recent criticisms coming from the cardiovascular research community (3, 4). Reexamining the role of AUC has been motivated in part by frustration at not being able to identify valuable biomarkers on the basis of AUC. The AUC is often interpreted as the probability of correct ordering—correct in the sense that when comparing the risk predictions for two subjects, only one of whom develops breast cancer, the risk calculated for the breast cancer subject is the larger of the two values. The first criticism of the AUC is that this probability is not a clinically relevant way of summarizing predictive performance. Subjects do not present to the clinic in pairs, and the problem is not to determine which of the pair will develop cancer. A second criticism is that the scale is somewhat deceptive. A substantial gain in performance may not result in a substantial increase in AUC. For example, suppose that the sensitivity changes from 40% to 90% but only over the range of specificities corresponding to 90% – 100%.

Correspondence to: Margaret S. Pepe, PhD, Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109 (mspepe@u.washington.edu).

This is an enormous improvement in performance: while maintaining specificity at 100%, now 90% rather than 40% of cancers can be predicted. However, the change in AUC is only 0.05. Although an extreme example, it illustrates the point. These two criticisms of AUC apply generally, not solely to risk prediction. The AUC really is a poor metric for evaluating markers for disease diagnosis, screening, or prognosis. The third criticism, which is specific to risk prediction, is that the AUC, and indeed the ROC curve itself, hides the values of risk calculated for subjects in the population. Indeed, the risk values are not visible from the ROC curve or the related curves in figure 2 of Gail (1). Moreover, the same ROC curve results if risk values are transformed monotonically, say, multiplied by a factor of 10, yet the clinical implications of these risk values would be very different.

The key question in evaluating risk prediction models concerns the number of subjects who are identified as being at high risk. Does a model that includes SNPs identify a substantially larger number of women at high risk for breast cancer who might therefore benefit from an intervention? In other words, does it do a better job than a model without SNPs at risk stratifying the population? The population distributions of absolute risk derived from the two models can be displayed to allow this sort of assessment. At any chosen high-risk threshold, the proportions of subjects with risks above that threshold can be compared. We refer to these plots as “predictiveness curves” (5, 6). Cook (3) tabulates these proportions for specific thresholds considered therapeutically relevant in preventing cardiovascular events. Unfortunately, it is not possible to determine the population distributions of absolute risk for the models described in Gail (1). The distributions of relative risk are shown instead. Gail notes that to derive the absolute risk distribution from the relative risk distribution, one needs to know the absolute risk in the baseline group, those with the lowest level of risk for all factors in the model, denoted by Gail with the letter *k*. The effect of *k* would be to shift the relative risk distribution shown in his figure 1 by log(*k*) to arrive at the distribution curve for absolute risk. Because the models differ in risk factors included, the baseline group varies across models and so too does the corresponding risk, *k*. This means that the curves in figure 1 would need to be shifted by different degrees to assemble the absolute risk distributions from them. In conclusion, the comparison of relative risk distributions does not give direct information about the comparison of absolute risk distributions, which is of key interest for comparing risk prediction models.

Interestingly, the absolute risk distributions could have been calculated by Gail if population incidence rates were specified. In particular, the absolute risk in the baseline group, *k*, for each model is simply the population incidence divided by the average relative risk. Because *k* is the factor that links the relative risk distribution to the absolute risk distribution, values for age-specific incidence of breast cancer could therefore be used in conjunction with Gail’s calculations to determine the age-specific risk distributions for women using models with and without SNPs. The age-specific proportions of women identified at high risk could then be compared across models. This would be an interesting exercise to complement Gail’s calculations.

Appropriate evaluation of risk prediction models requires specification of a risk threshold for defining individuals as high risk. What high-risk threshold should be used in the breast cancer setting? A consensus on this fundamental question does not exist at present. The choice depends on costs and benefits associated with interventions that will be employed for women designated as high risk. Tamoxifen therapy and screening with magnetic resonance imaging are among the set of options for breast cancer. Medical decision theory provides an explicit solution for high-risk designation in terms of 1) the net benefit, *B*, of being classified as high risk if, in the absence of intervention, one is destined to develop breast cancer, and 2) the net cost, *C*, of being classified as high risk if, in the absence of intervention, one is destined not to develop breast cancer. The risk threshold at which expected benefit exceeds

expected cost is $C/(C + B)$ (7). The higher the cost:benefit ratio, the higher the optimal threshold. Cardiovascular consensus groups (8) have determined risk thresholds based on costs and benefits of different therapy options. Corresponding guidelines for defining high risk in the context of breast cancer prevention must be developed. We need them to gauge the value of risk prediction models. Risk thresholds might be chosen to vary with factors such as age, recognizing that costs and benefits of high-risk interventions are not uniform across the population. Moreover, in practice each woman may have her own tolerance for risk that could vary from guidelines developed by consensus groups.

Gail's ROC analysis indicates that even under optimistic assumptions, SNPs—or, for that matter, other risk factors with moderate relative risks—are unlikely to substantially improve current algorithms for breast cancer risk prediction. The same conclusion may well hold with analyses that focus on proportions of high-risk (or low-risk) women identified. Indeed, this has been the experience in cardiovascular research. Biomarkers such as C-reactive protein (CRP) and high-density lipoproteins that do not increase AUC statistics do not appear to improve risk stratification either, at least when considering the population as a whole (9). Subsets of the population may, however, benefit from information in these markers. Ridker and Cook (10) report that for subjects at intermediate risk according to standard risk factors, CRP can further stratify a large fraction of subjects into high- and low-risk categories. Similarly, for breast cancer risk prediction SNPs and biomarkers may have their greatest impact on subpopulations.

Risk stratification is not the only component of prediction model evaluation. As Gail notes, calibration is of paramount importance. A well-calibrated model ensures that the calculated risks reflect the actual proportions of subjects who develop disease. Also of interest is the accuracy of risk classifications, defined as the proportion of women who develop breast cancer who are classified as high risk and the proportion of women who do not develop breast cancer who are classified as low risk (5). These can also be calculated using the absolute risk distribution, a fact previously noted by Gail and Pfeiffer (11).

Acknowledgments

Funding

National Institutes of Health (GM054438 to M.S.P., CA086368 to M.S.P. and H.E.J.).

References

1. Gail M. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst. 2008; 100(14):1037–1041. [PubMed: 18612136]
2. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol. 1975; 12(6):387–415.
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115(7):928–935. [PubMed: 17309939]
4. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008; 27(2):157–172. discussion 207 – 212. [PubMed: 17569110]
5. Pepe MS, Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008; 167(3):362–368. [PubMed: 17982157]
6. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics. 2007; 63(4):1181–1188. [PubMed: 17489968]
7. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. N Engl J Med. 1975; 293(5):229–234. [PubMed: 1143303]

8. Expert Panel on Detection EaToHBCiA. Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). J Am Stat Assoc. 2001; 285(19):2486–2497.

9. Pepe MS, Janes H, Gu JW. Letter by Pepe et al regarding article, “Use and misuse of the receiver operating characteristic curve in risk prediction.”. Circulation. 2007; 116(6):e132. author reply e134. [PubMed: 17679623]

10. Ridker PM, Cook N. Clinical usefulness of very high and very low levels of C-reactive protein across the full range of Framingham risk scores. Circulation. 2004; 109(16):1955–1959. [PubMed: 15051634]

11. Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics. 2005; 6(2): 227–239. [PubMed: 15772102]