In their “Discussion,” Gail et al. (1) seem to be saying that a problem with a model’s concordance statistic (or area under a receiver operating curve [AUC]) is that it is inherently retrospective, implying that one should therefore not place much stock in it. The concordance statistic from the Gail model is retrospective in the same sense that the Gail model’s parameter estimates (which produce the estimated risks) are retrospective; that is, all of these measures derive from existing data on risk factors and outcomes. As is the case with the model’s parameter estimates, if the model is applied to a sample or population similar in key respects to that from which it has been derived, we can reasonably expect that the AUC will be generalizable. The key issue about the AUC is not, then, that it is retrospective or at least any more retrospective than parameter estimates.

The authors next state that “…if a model includes age only and predicts that every woman in the age range 60-64 years has a 1.7% risk of breast cancer … then the AUC will be 0.50. Some would mistakenly construe this to mean that the model does not perform any better than a coin flip in predicting who will or will not get breast cancer. In fact, if one predicted that none of these women would be diagnosed with invasive breast cancer in the next five years, one would predict correctly for 100 – 1.7 = 98.3% of the women.” This latter statement is true, but one does not need to rely on such a statistical model as support for a prediction that no women will get breast cancer during a period of 5 years. One simply needs general population incidence rates to see that only a small minority of women will develop disease over a period of 5 years. In recent years, the Gail model has been promoted as a decision-making aid for women with respect to chemoprevention of breast cancer. If the goal in providing risk estimates is to help women with such dichotomous decision making about prophylaxis, or even with entrance (or not) into a clinical trial, one cannot imply that a model is doing a good job by giving everyone the same risk estimate. A model that gives each person the same risk estimate has zero discriminatory ability—no individual appears better off or worse off than anyone else—which is precisely what a concordance statistic of 0.5 means in this example. Along these lines, the observed concordance statistic for the Gail model in African American women aged 60–64 years, 0.507, is indicative of no discriminatory ability. This result is troubling given that the age group of 60–64 years is the 5-year age group, of those examined by the authors, with the highest incidence.

Gail et al. (1) imply a key question in their “Discussion”: might it be better, overall, to tell all women in the general population that they will not get breast cancer in the next 5 years and to be wrong only a small percentage of the time? (I am excluding from consideration here women known to be at high genetic risk—the Gail model is not intended for such women anyway.) According to the provocative work of Tu et al. (2), it is possible to lose information (i.e., to increase entropy) by administering a
screening test and by then attempting to segregate individuals into two groups. In other words, one could make mistakes in a higher percentage of the total population, and mistakes of a more harmful nature, by attempting to segregate individuals into two groups (eg, high and low risk) rather than by not segregating in the first place.

On the basis of the work of Tu et al. (2), it is likely that there will be a loss of information, and a consequent high risk of error, when attempts are made to segregate individuals into high-risk and low-risk groups by use of models with low discriminatory accuracy. It would be helpful if the authors could speak more on the intended and unintended uses of their models with respect to such dichotomous segregation of women.

BEVERLY ROCKHILL LEVINE

References


Notes

Correspondence to: Beverly Rockhill Levine, PhD, Department of Public Health Education, 437K HHP Bldg, University of North Carolina, Greensboro, NC 27403 (e-mail: b llevine@uncg.edu).

DOI: 10.1093/jnci/djn106 © The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oxfordjournals.org.

Response

We wrote (1) that the area under the curve (AUC) is inherently a retrospective quantity because it can be estimated from a random sample of case patients and of control subjects. Arguably better criteria for assessing risk models (2–4), such as positive and negative predictive value, misclassification rate, calibration, and expected loss in a decision problem, cannot be estimated from case patients and control subjects; follow-up data on the probability of disease are needed.

A model that is based solely on age-specific incidence rates can be useful, even though it has an age-specific AUC of 0.5. If a 62-year-old white woman knows, on the basis of her age alone, that her chance of developing breast cancer is 1.7% in the next 5 years, she should carefully consider whether to take tamoxifen to prevent breast cancer, because the tamoxifen-induced increases in the risks of stroke and endometrial cancer exceed the decrease in breast cancer risk, as shown in table 10 of Gail et al. (5).

We stated (1): “A limitation of the CARE model is that it has low age-specific discriminatory accuracy as measured by the concordance or AUC...” where the Women’s CARE Study is the Women’s Contraceptive and Reproductive Experiences Study. Levine points to the AUC of 0.507 for African American women who are aged 60–64 years. The average age-specific AUC, 0.555 (95% confidence interval [CI] = 0.535 to 0.575), measures the discriminatory ability of the risk factors for women in a given age group. If, as is common in the cardiovascular literature, one also credits age when computing the AUC, the CARE model AUC (1) increases to 0.636 (95% CI = 0.617 to 0.655).

Levine states: “Gail et al. imply a key question in their discussion: might it be better, overall, to tell all women in the general population they will not get breast cancer in the next 5 years and be wrong only a small percentage of the time?” We certainly did not mean to imply this question. Indeed, we disagree with the idea that one should use an estimated 5-year risk to decide whether a woman will or will not develop breast cancer. Rather, that estimate should provide perspective on the level of risk and allow the woman to compare that risk with other risks that she might face. A formal analysis of losses from misclassification (3) indicates that high discriminatory accuracy is needed to reliably screen the general population to identify women for whom special diagnostic or other measures are indicated. However, well-calibrated models with modest discriminatory accuracy can be useful for deciding whether an intervention, such as tamoxifen, that has offsetting risks and benefits should be used.

Levine asserts that it is possible to “increase entropy” by “attempting to segregate individuals ...into ...high- and low-risk” groups. Suppose r(X) is a well-calibrated (3) risk model for a person with risk factors X. The average risk is \( \bar{r} = \int r(x) dF(x) \), where \( F \) is the distribution of X. The entropy, given X, is \( H(r(x)) = -r(x) \log r(x) - (1-r(x)) \log(1-r(x)) \). Because \( H \) is concave, Jensen’s (6) inequality implies that \( \int H(r(x)) dF(x) \leq H(\bar{r}) \). This result is true for any set of risk factors and joint distribution \( F \). Thus, a well-calibrated model, such as the CARE model, will never increase entropy above that which results from assigning all women the average risk. Likewise, a well-calibrated model that partitions the population into high- and low-risk groups cannot increase entropy, contrary to Levine’s assertion.

MITCHELL H. GAIL
Leslie Bernstein
JOSEPH P. COSTANTINO
DAVID PEE
GISKE URSIN

References


