Evaluating Prognostic Markers Using Relative Utility Curves and Test Tradeoffs

Stuart G. Baker and Barnett S. Kramer

All authors: National Cancer Institute, Bethesda, MD.

Published online ahead of print at www.jco.org on June 29, 2015.

Editor's note: Statistics in Brief articles are short communications regarding statistical methods or issues. They are designed to alert and educate the readership about a method or issue that may be unfamiliar to or underused by the clinical research community. Each article will serve as a short primer and may refer the reader to additional sources for detailed information regarding both background and application.

Supported by the Division of Cancer Prevention of the National Cancer Institute.

The funder did not have a role in the writing of the article or the decision to submit the article for publication. The opinions presented here should not be viewed as official opinions or positions of the National Cancer Institute.

Authors' disclosures of potential conflicts of interest are found in the article online at www.jco.org. Author contributions are found at the end of this article.

Corresponding author: Stuart G. Baker, ScD, 9609 Medical Center Dr, MSC 9789, Bethesda MD 20892-9789; e-mail: sb16i@nih.gov.

Published by the American Society of Clinical Oncology

0732-183X/15/3323w-2578w/\$20.00

DOI: 10.1200/JCO.2014.58.0092

Markers for predicting the risk of cancer incidence or recurrence in individuals are called prognostic markers.¹⁻³ In contrast, markers for predicting the effect of treatment on outcome are called predictive markers.¹⁻³ Our focus is on evaluating prognostic markers; evaluating predictive markers involves a different type of analysis.⁴ In clinical cancer research, there is often a standard risk-prediction model for predicting cancer incidence or recurrence. Let us call this model 1. Depending on the context, investigators are interested in identifying either persons at high risk or persons at low risk under model 1; these persons would receive an intervention that differs from the default management in the absence of risk prediction. Often, investigators would like to improve clinical decision making by fitting a risk prediction model based on all the variables in model 1 plus an additional prognostic marker or set of prognostic markers. Let us call this model 2. Just as with model 1, persons at high (or low) risk under model 2 would receive an intervention that differs from the default management in the absence of risk prediction. The basic question is whether any improvement in risk prediction with model 2 instead of model 1 is large enough to justify collection of data on the additional prognostic marker(s) in model 2. This question is particularly relevant when data collection is expensive or involves an invasive procedure.

In our example involving the prediction of invasive breast cancer in asymptomatic women,⁵ model 1 includes age, family history, variables related to personal reproductive history, and personal medical history related to cancer; model 2 includes all the variables in model 1 plus a set of single-nucleotide polymorphism (SNP) markers; the goal is to identify high-risk women to receive chemoprevention. As another example, investigators may add markers to standard variables predicting postresection residual tumors in patients with early-stage testicular cancer, where the goal is to identify patients at low risk for whom surveillance is preferable to adjuvant chemotherapy.^{6,7} Other examples include adding breast density to standard variables predicting the incidence of invasive breast cancer⁸ and adding markers to smoking duration when predicting lung cancer incidence.⁹

Investigators fit models 1 and 2 to data from what is called a training sample. Evaluating the performance

of the models in the training sample can be problematic because of overfitting, which means the model fit is overly sensitive to chance deviations. Ideally, investigators should evaluate the performances of the models in an independent test sample from a different population than that associated with the training sample.^{10,11} For comparing the performances of models 1 and 2, three popular statistical measures are: (1) the coefficient for the marker in a regression model (eg, odds ratio) in model 2 compared with model 1,¹² (2) the change in area under the receiver operating characteristic (ROC) curve (AUC) between models 1 and 2,^{13,14} and (3) net reclassification improvement comparing models 1 and 2.¹⁵ For the example involving SNPs, the odds ratios for SNPs ranged from 1.08 to 1.35, the AUC increased from 0.557 to 0.594, and net reclassification improvement equaled 0.195.⁵ On the basis of these changes in the odds ratio or the net reclassification improvement, some investigators may conclude that model 2 substantially improves risk prediction. In contrast, on the basis of the change in the AUC, some investigators may conclude the opposite, namely that model 2 leads to little improvement in risk prediction. It is not possible to say which of these conclusions is most appropriate. The fundamental problem is that these purely statistical approaches cannot answer the basic question about the value of model 2 versus model 1 in terms of clinical decision making.

To address the basic question, we use a decision-analytic approach, the premises of which are: (1) risk prediction is only valuable to the extent that it improves clinical decision making, and (2) the value of clinical decision making depends on the costs and benefits of these decisions. To avoid detailed specifications of costs and benefits, one can use either decision curves^{16,17} or related relative utility curves.^{8,18} Our focus is on relative utility curves, which we explain in terms of the SNP example. Relative utility curves involve the following five fundamental quantities. The prevalence (P) is the probability of developing invasive breast cancer in the population of asymptomatic women targeted by the trial. The true-positive rate at cut point R (TPR_R) is the probability of a correct prediction of invasive cancer at values of the predicted risk of invasive breast cancer greater than R . The false-positive rate at cut point R (FPR_R) is the probability of an incorrect prediction

Fig 1. (A) Receiver operating characteristic and (B) relative utility (RU) curves for predicting risk of invasive cancer in asymptomatic women. Model 1 includes standard risk factors, and model 2 adds single-nucleotide polymorphisms (SNPs). Vertical dashed lines shown with RU curves indicate risk thresholds where test tradeoff is computed. AUC, area under receiver operating characteristic curve; FPR, false-positive rate; TPR, true-positive rate.

of invasive cancer at values of the predicted risk of invasive cancer greater than R . The benefit of a true positive (B) is the anticipated reduction in the risk of invasive breast cancer as a result of chemoprevention. The cost of a false positive (C) is the anticipated cost, in terms of adverse effects and monetary costs, of administering chemoprevention in persons who would not develop invasive breast cancer in the absence of chemoprevention.

To derive the formula for relative utility, we need to consider costs and benefits in the same units. However, to use relative utilities, clinicians do not need to explicitly specify these quantities. The net benefit of prediction is the anticipated benefit of a true positive minus the anticipated cost of a false positive, namely $NB_{\text{prediction}} = B \times P \times TPR_R - C \times (1 - P) \times FPR_R$. The net benefit of perfect prediction is the net benefit of prediction in the ideal case when $TPR_R = 1$ and $FPR_R = 0$, namely $NB_{\text{perfect}} = P \times B$. The ratio of these net benefits is relative utility, namely $RU = NB_{\text{prediction}} / NB_{\text{perfect}} = TPR_R - FPR_R \{ (1 - P) / P \} \times C / B$. The rationale for relative utility now becomes clear: It is a meaningful function of net benefit that depends on the ratio C/B . Because relevant values of C/B are difficult to specify, it is preferable to rewrite the relative utility using the risk threshold: $T = (C/B) / \{ 1 + (C/B) \}$. On the basis of decision analysis,^{18,19} a person with risk threshold T is indifferent between receiving chemoprevention or no chemoprevention after taking into account all the costs and benefits. With a little mathematic finesse, one can write the relative utility in standard form: $RU(T) = TPR_T - FPR_T \{ (1 - P) / P \} \{ T / (1 - T) \}$. The relative utility curve plots $RU(T)$ versus T . Figure 1 also shows the corresponding ROC curve. Investigators need to identify a relevant range of risk thresholds. On the basis of decision analysis,¹⁸ the default of no chemoprevention in the absence of risk prediction implies that the lower bound of the relevant range of risk thresholds is P . The upper bound should occur at a risk threshold less than where the relative utility curve for model 2 approaches zero.

At various risk thresholds within the relevant range, investigators compute the test tradeoff as a function of P and the difference in relative

utility curves for models 1 and 2. The test tradeoff is a break-even point for the overall net benefit, which is the net benefit minus the cost of collecting data on the additional prognostic marker in model 2.¹⁸ For risk thresholds in the relevant range, the test tradeoff for the addition of SNPs to the risk prediction model had a maximum value of approximately 3,100. A test tradeoff of 3,100 implies that SNP data must be collected from at least 3,100 persons for every correct prediction of invasive breast cancer to increase the overall net benefit (and hence be worthwhile). Given the small expense and harm of collecting SNP data, a test tradeoff of 3,100 may be reasonable. Although the interpretation of the test tradeoff involves a subjective component, it is more clinically informative than an odds ratio, change in AUC, or net reclassification improvement and thus may be more useful to clinicians.

In summary, we recommend that investigators compute relative utility curves and test tradeoffs for any evaluation of prognostic markers in a risk prediction model. If the ROC curves are monotonically increasing, investigators can apply simple formulas for risk intervals.^{8,18} Otherwise, investigators can either reduce the number of risk intervals to create monotonically increasing ROC curves and then apply the simple formulas, or use software²⁰ written in Mathematica (version 10.0; Wolfram Research, Champaign, IL) that optimally creates monotonically increasing ROC curves and computes relative utility curves and test tradeoffs.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at www.jco.org.

AUTHOR CONTRIBUTIONS

Manuscript writing: All authors

Final approval of manuscript: All authors

REFERENCES

1. Sargent DJ, Mandrekas SJ: Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. *Clin Trials* 10:647-652, 2013
2. Sargent DJ, Conley BA, Allegra C, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 23:2020-2027, 2005
3. Italiano A: Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol* 29:4718, 2011
4. Baker SG, Kramer BS, Sargent DJ, et al: Biomarkers, subgroup evaluation, and trial design. *Disc Med* 13:187-192, 2012
5. Mealiffe ME, Stokowski RP, Rhee BK, et al: Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J Natl Cancer Inst* 102:1618-1627, 2010
6. Baker SG, Van Calster B, Steyerberg EW: Evaluating a new marker for risk prediction using the test tradeoff: An update. *Int J Biostat* 22:8, 2012
7. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 21:128-138, 2010
8. Baker SG: Putting risk prediction in perspective: Relative utility curves. *J Natl Cancer Inst* 101:1538-1542, 2009
9. Raji OY, Duffy SW, Agbaje OF, et al: Predictive accuracy of the liverpool lung project risk model for stratifying patients for computed tomography screening for lung cancer: A case-control and cohort validation study. *Ann Intern Med* 157:242-250, 2012
10. Collins GS, de Groot JA, Dutton S, et al: External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 14:40, 2014
11. Schumacher M: Probability estimation and machine learning. *Biom J* 56:531-533, 2014
12. Spitz MR, Amos CI, D'Amelio A Jr, et al: Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst* 101:1731-1732, 2009
13. Tzoulaki I, Liberopoulos G, Ioannidis JP: Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 302:2345-2352, 2009
14. Pepe MS, Janes H, Longton G, et al: Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 159:882-890, 2004
15. Leening MJ, Vedder MM, Witteman JC, et al: Net reclassification improvement: Computation, interpretation, and controversies: A literature review and clinician's guide. *Ann Intern Med* 160:122-131, 2014
16. Vickers AJ, Elkin EB: Decision curve analysis: A novel method for evaluating prediction models. *Med Dec Making* 26:565-574, 2006
17. Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Dec Mak* 8:53, 2008
18. Baker SG, Schuit E, Steyerberg EW, et al: How to interpret a small increase in AUC with an additional risk prediction marker: Decision analysis comes through. *Stat Med* 33:3946-3959, 2014
19. Pauker SG, Kassirer JP: Therapeutic decision making: A cost-benefit analysis. *N Engl J Med* 293:229-234, 1975
20. National Cancer Institute: Evaluating risk prediction markers for binary outcomes via relative utility curves. <http://prevention.cancer.gov/about-dcp/staff-search/stuart-g-baker-scd/evaluating-risk-prediction>

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Evaluating Prognostic Markers Using Relative Utility Curves and Test Tradeoffs

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or jco.ascopubs.org/site/ifc.

Stuart G. Baker

No relationship to disclose

Barnett S. Kramer

No relationship to disclose