HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2018 December 10.

Published in final edited form as:

Stat Med. 2017 December 10; 36(28): 4491–4494. doi:10.1002/sim.7132.

The summary test tradeoff: a new measure of the value of an additional risk prediction marker

Stuart G. Baker[*]

[*]Biometry Research Group, National Cancer Institute, Bethesda, *M*D 20892 USA. sb16i@nih.gov

Keywords

prognostic marker; relative utility curve; risk prediction; ROC curve; Youden index

For evaluating differences in classification performance between risk prediction models with and without an additional marker, Pencina, Steyerberg, and D'Agostino, Sr [1] proposed as a summary measure the net reclassification improvement at the event rate. To support their proposal, they showed that, under perfect calibration, the net reclassification improvement at the event rate equals both the difference in maximum Youden indexes and the difference in maximum relative utilities. I extend their viewpoint by proposing a new measure called the summary test tradeoff. Although the summary test tradeoff is a function of the difference in maximum relative utilities, it is more informative for decision-making than the difference in maximum relative utilities. Another advantage of the summary test tradeoff is that it does not require perfect calibration for valid estimates. Before discussing the interpretation and use of the summary test tradeoff, I discuss its derivation by briefly reviewing the theory underlying relative utility curves.

Summarizing the risk profile

Let the target sample denote a sample from a target population for which the risk prediction models have the same Receiver Operating Characteristic (ROC) curves as in the target population. The probability of event can differ between the target sample and the target population, as would arise if investigators created the target sample by a separate random sampling from persons with and without the event in the target population.

Let x denote the possibly multidimensional risk profile of a person in the target sample. Let $score(x)$ denote a scalar measure summarizing the risk profile x. A simple example is $score(x)$ equal to x where is x is a single marker such as cholesterol level. A complex example is $score(x)$ equal to the difference between distances to centroids for two classification groups, using centroids from a separate training sample. Pencini et al. [1] consider the special case in which $score(x)$ is the estimated risk (probability of the event) based on model fit to a separate training sample, a quantity they denote $r(x)$. Pencini et al. [1] assume perfect calibration, which says that $r(x)$ equals the risk of the event for a person in the target population with risk profile x.

The ROC curve and its slope

Let t denote a threshold for $score(x)$. Let $TPR(t) = pr(score(x) \geq t \mid event)$ denote the true positive rate, or sensitivity, at threshold t. Let $FPR(t) = pr(score(x) \geq t \mid no\ event)$ denote the false positive rate, or 1–specificity, at threshold t. The ROC curve is a plot of $TPR(t)$ versus $FPR(t)$. Let $R(t)$ denote the risk (probability of event) in the target population at threshold t of the score. Let p denote the probability of event in the target population. Writing $R(j) = pr(event \mid score = j)$, $W(j) = pr(score = j)$, and $p = pr(event)$ gives

$$\begin{aligned} TPR(t) &= \int_t^{\infty} pr(event|score=j)\ pr(score=j)\ dj/pr(event) \\ &= \int_t^{\infty} R(j)\ W(j)\ dj/p, \end{aligned} \quad (1)$$

$$\begin{aligned} FPR(t) = & \int_t^{\infty} pr(no\ event|score=j)\ pr(score=j)\ dj/pr(no\ event) \\ & = \int_t^{\infty} \{1 - R(j)\}\ W(j)\ dj/(1-p). \end{aligned} \quad (2)$$

Therefore, $dTPR/dt = -R(t)\ W(t)/p$ and $dFPR/dt = -(1–R(t)\}\ W(t)/(1–p)$. These equations yield the following important relationship, which was previously derived for discrete ROC curves [2],

$$\begin{aligned} ROCSLOPE(t) &= dTPR/dFPR = (dTPR/dt)/(dFPR/dt) \\ &= \{(1-p)/p\} \quad \{R(t)/(1-R(t)\}. \end{aligned} \quad (3)$$

Relative utility

A net benefit is the total of benefits minus harms. Each person has a set of four utilities, which are associated with false positive, true positive, false negative, and true negative outcomes. The sum of these utilities weighted by the probabilities of the corresponding outcomes yields the net benefit of risk prediction [2]. A parsimonious function of these utilities is the risk threshold, denoted by T, which is the risk at which a person would be indifferent between treatment and no treatment [2, 3]. The motivation for the relative utility is that it an easily interpretable quantity that is a function of the four utilities only through T. The relative utility has the general form,

$$RU(T) = \{max_t NB_{Pred}(t) - NB_{NoPred}\} / \left\{NB_{PerfectPred} - NB_{NoPred}\right\}. \quad (4)$$

where $NB_{Pred}(t)$ = net benefit of risk prediction at threshold t for the score(x),

$$\begin{aligned} NB_{PerfectPred} &= \text{net benefit of perfect prediction,} \\ NB_{NoPred} &= \text{maximum net benefit of no risk prediction.} \end{aligned}$$

In other words, the relative utility is the maximum net benefit of risk prediction (where the maximum is over all possible thresholds for the score) in excess over the net benefit of no prediction relative to the net benefit of perfect prediction in excess over the net benefit of no prediction. A reasonable requirement for good risk prediction is that $R(t)$ increases as t increases, which implies a concave ROC curve. If this requirement does not hold, it is possible to redefine $R(t)$ by creating a concave ROC curve from a non-concave ROC curve [2]. With this requirement, the relative utility simplifies to

$$RU(T) = \{NB_{Pred}(t) - NB_{NoPred}\} / \left\{NB_{PerfectPred} - NB_{NoPred}\right\}, \quad \text{where } t \text{ solves the equation } R(t) = T. \tag{5}$$

The maximum net benefit of no prediction is the larger of two quantities: (*i*) the net benefit of no prediction with no treatment and (*ii*) the net benefit of no prediction with treatment. Applying these two cases with extensive simplification yields the standard formula for relative utility [2],

$$RU(T) = \begin{cases} \{1 - FPR(t)\} - \{1 - TPR(t)\} / ROCSLOPE(t), & \text{if } T < p, \\ \{TPR(t) - ROCSLOPE(t)\, FPR(t)\}, & \text{if } T \geq p, \end{cases} \quad \text{where } t \text{ solves the equation } R(t) = T. \tag{6}$$

Pencini et al. [1] presented a special case of the relative utility in equation (6) for perfect calibration of the score, which corresponds to $R(t) = t$. The reason the relative utility in equation (6) is appropriate without perfect calibration of the score is that the risk thresholds apply to the risks in the target population, which is, in a sense, a built-in perfect calibration. A relative utility curve is a plot of $RU(T)$ versus T. The maximum of the relative utility curve is $maxRU = RU(p)$. At this maximum, the slope of the ROC curve equals 1. For $T > p$, $RU(T)$ monotonically decreases to 0 as T increases, and for $T < p$, $RU(T)$ monotonically decreases to 0 as T decreases.

Maximum Youden index

The Youden index is $YI(t) = TPR(t) - FPR(t)$. When $R(t)$ increases a t increases, so the ROC curve is concave, the maximum of the Youden index occurs at threshold t such that $dYI(t)/dt = dTPR(t)/dt - dFPR(t)/dt = 0$, which implies $ROCSLOPE(t) = dTPR(t) / dFPR(t) = (dTPR(t)/dt) / dFPR(t)/dt = 1$. Therefore, $maxYI = YI(t_{ROCslope=1})$ where $t_{ROCslope=1}$ is the threshold where the slope of the ROC curve equals 1.

Summary test tradeoff

Let subscript “new” refers to the risk prediction model with the additional marker and subscript “old” refer to the risk prediction model without the addditonal marker. The test tradeoff is $TestTradeoff(T) = 1 / [\{p\{RU_{new}(T) - RU_{old}(T)\}]$. The test tradeoff is the minimum number of tests for a new marker that need to be traded for a true positive to yield a positive

net benefit [2]. The summary test tradeoff equals the test tradeoff when the risk threshold equals the event rate. It can be written in the following equivalent formulas,

$$\begin{aligned} Summary\ TestTradeoff = &\ Text\ Tradeoff\ (p) \\ = &\ 1/\left[p\ \{RU_{new}\,(p) - RU_{old}\,(p)\}\right] \\ = &\ 1/\left\{p\,(maxRU_{new} - maxRU_{old})\right\} \\ = &\ 1/\left\{p\,(maxYI_{new} - maxYI_{old})\right\} \\ = &\ 1/\left[p\left\{YI_{new}\,(t_{ROCSLOPEnew=1}) - YI_{old}\,(t_{ROCSLOPEold=1})\right\}\right]. \end{aligned} \quad (7)$$

The last formula in equation (7) yields a simple method for computing the summary test tradeoff, namely, for each risk prediciton model, computing the threshold at which the slope of the ROC curve equals 1 and subsituting this threshold into the Youden index.

Because relative utility curves montonically decrease as T gets farther from $T = p$, the difference in relative utility curves at $T = p$ is usually the largest or close to the largest difference in relative utilities for a given T. Consquently, the summary test tradeoff is the smallest or close to the smallest test tradeoff. Therefore, the summary test tradeoff provides a reasonable summary measure to gauge the worth of ascertaining an additional marker to include in the risk prediction model, by providing an approximate (or exact) lower bound for the test tradeoff.

As an example, consider a summary test tradeoff of 3000 for including a new marker for predicting the risk of cancer. In this example, the summary test tradeoff is the minimum number of ascertainments of the new marker that would be traded for a correct prediction of cancer incidence to yield a positive net benefit. If ascertaining the new marker involves an invasive test with a significant risk of mortality, the summary test tradeoff indicates that the marker would not be worthwhile for adding to a risk prediction model. Even if the summary test tradeoff was an approximate and not an exact lower bound on the test tradeoff, the conclusion would remain the same for other risk thresholds. In contrast, if ascertaining the new marker involved answering a questionnaire, the summary test tradeoff suggests that including the new marker is worthwhile. In this case, even if the summary test tradeoff were considerably smaller than the test tradeoff at a more relevant risk threshold, the conclusion for the more relevant risk threshold would likely be the same.

Although a formal sensitivity analysis involving test tradeoffs at different risk thresholds would be more informative than a single measure, the summary test tradeoff is easier to compute and report. A problem with a full sensitivity analysis is that test tradeoffs at extreme risk thresholds can be unreliable due to small sample sizes and convergence of relative utilities toward zero. If there is interest in a single summary measure of the performance of an additional marker for risk prediction, the summary test tradeoff should be considered.

Acknowledgments

This work was supported by the National Cancer Institute.

References

1. Pencina MJ, Steyerberg EW, D'Agostino RB Sr. Net reclassification index at event rate: properties and relationships. Statistics in Medicine. Version of Record online. 18 JUL 2016.
2. Baker SG, Van Calster B, Steyerberg EW. Evaluating a new marker for risk prediction using the test tradeoff: An update. International Journal of Biostatistics. 2012; 8:5.
3. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. New England Journal of Medicine. 1975; 293(5):229–34. [PubMed: 1143303]