NIH Public Access
Author Manuscript
Clin Trials. Author manuscript; available in PMC 2014 April 01.

Published in final edited form as:
Clin Trials. 2013 October ; 10(5): . doi:10.1177/1740774513496490.

Testing the Incremental Predictive Accuracy of New Markers

Colin B. Begg, **Mithat Gonen**, and **Venkatraman E. Seshan**
Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Abstract

Background—It has become commonplace to use receiver operating curve (ROC) methodology to evaluate the incremental predictive accuracy of new markers in the presence of existing predictors. However, concerns have been raised about the validity of this practice. We have evaluated this issue in detail.

Results—Simulations have been used that show clearly that use of risk predictors from nested models as data in subsequent tests comparing areas under the ROC curves of the models leads to grossly invalid inferences. Careful examination of the issue reveals two major problems: (1) the data elements are strongly correlated from case to case; and (2) the model that includes the additional marker has a tendency to interpret predictive contributions as positive information regardless of whether observed effect of the marker is negative or positive. Both of these phenomena lead to profound bias in the test.

Conclusions—We recommend strongly against the use of ROC methods derived from risk predictors from nested regression models to test the incremental information of a new marker.

Introduction

In evaluating the incremental predictive or diagnostic accuracy of a new marker it is commonplace to test the contribution of the marker in a nested regression model that includes known predictors and then, if this test is significant, to use a further test comparing the areas under the receiver operating characteristic (ROC) curves derived from the baseline model and from the model augmented with the new marker. Not only is this strategy logically inconsistent, since it is essentially the same null hypothesis that is being tested in both cases, but the test comparing the areas under the curves (AUC test) is often observed to be non-significant following a significant Wald test from the regression model. In earlier work by our group Vickers et al. demonstrated by simulation that the AUC test is exceptionally conservative in this setting, with very low power and test size far below the nominal level.[1] The purpose of this investigation was to follow-up the findings of Vickers et al. to determine the reasons for the inadequacies of the AUC test in this context. Throughout we used the AUC test proposed by DeLong et al.[2] This widely-used test has been shown to have valid statistical properties in simulations comparing separate diagnostic tests where the results of individual patients are independent though the tests results are correlated, the setting for which this test was originally developed.[3]

Author for correspondence: Colin B. Begg, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 307 East 63rd St, New York, NY 10065. Phone: (646) 735-8108, Fax: (646) 735-0009, beggc@mskcc.org.

Conflict of Interest Statement
The authors declare that there is no conflict of interest.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Results

In the context of assessing predictive accuracy using regression models the AUC test employs risk predictors for each patient derived from each of the two nested regression models. These are then used as data elements in the test. Unfortunately this strategy seriously violates the assumptions embedded in the AUC test framework, namely that the sets of observations from each patient are i.i.d., and that under the null hypothesis that the marker possesses no incremental diagnostic information the baseline and the expanded models have the same AUC. The first of these problems concerns the absence of independence of results from different patients. Risk predictors are functions of the estimated regression parameters and the factors included in the model. As a result risk predictors from patients with similar risk factor profiles are correlated, in some cases strongly correlated. The averages of these correlations are frequently as high as 0.5 even in typical data scenarios. The second problem is a little more subtle, concerning a phenomenon we term “known directionality”. When comparing two distinct diagnostic tests, the fundamental framework for which the AUC test was derived, the observed difference in AUCs under the null hypothesis is just as likely to favor one test as the other. However, this is not the case when comparing AUCs derived from predictors from nested models. When a null new marker is added to an existing model it is just as likely to lead to a negative coefficient in the regression as a positive one. However, the ROC curve is derived from the ranking of the resulting predictors, and these rankings typically lead to an apparent gain in predictive information, and thus a larger AUC. In other words the AUC modeling typically interprets a non-zero regression coefficient as contributing positive predictive information regardless of whether it is negative or positive. As a consequence, the AUC from the augmented model will usually, though not always, have a higher value than the AUC from the baseline model. This leads to a substantial bias in the mean of the AUC test statistic.

These two issues have a devastating effect on the validity of the AUC test. The distribution of the AUC test statistic has a mean value with a substantial positive bias under the null. Furthermore, the asymptotic variance substantially overestimates the true variance of the test statistic. As a result, despite the positive bias in the difference in AUCs the test fails to reach significance far more often than it should. This occurs both at the null and when the marker truly has incremental information, leading to a test with very low power and a true false positive rate that is far less than the nominal significance level of the test.

Despite the insensitivity of the AUC test in this context when compared with the Wald test of the regression coefficient the difference in the AUCs is not an inherently insensitive measure. This can be demonstrated in the following way. First, we can correct the biases in the asymptotic reference distribution to account for the induced correlations and the effects of known directionality. This can be accomplished by creating an orthogonal decomposition of the vector of the new marker values to create a vector of components that is orthogonal to the information from the baseline predictors. These orthogonal components can be permuted to create a new set of marker values and the regression parameters re-estimated to create a new test statistic. By repeating this process many times it is possible to generate the correct null distribution. Simulations show that this permuted AUC test has operating characteristics very similar to the Wald test with a modest loss of power. While we do not recommend this as an alternative to the Wald test, the results indicate that the difference in AUCs is an appropriate, sensitive measure for characterizing the impact of the new marker. Indeed other “measures” of predictive impact have been advocated in recent years, including the net reclassification improvement (NRI), a measure that reflects the extent to which the new predictive rule improves the classification of patients into clinically distinct categories, and the integrated discrimination improvement (IDI), an average of the improvements in sensitivity and specificity due to the new marker.[4] Both of these measures are biased in the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

same manner as the AUC difference due to the effects of known directionality and their sampling variances are similarly influenced by the lack of independence between the predictors from patient to patient.

We do not recommend the use of the AUC test to compare distinct markers from non-nested regression models. The problems due to dependence between patients' predictors remain, and although the test is less conservative than its nested counterpart the reference distribution is invalid. Problems also exist in using predictors to estimate parameters such as the difference in AUCs in independent "validation" samples, though the biases in this setting are modest.

Acknowledgments

This is a summary of a conference presentation for which most of the material presented is published in an article by the current authors already in print in Statistics in Medicine.[5] We appreciate that journal's policy that "If you wish to reuse your own article (or an amended version of it) in a new publication of which you are the author, editor or co-editor, prior permission is not required (with the usual acknowledgements)

Funding

This work was supported by the National Cancer Institute [grant numbers CA163251 andCA136783].

References

1. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Med Res Methodol. 2011; 11:13. [PubMed: 21276237]
2. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988; 44:837–845. [PubMed: 3203132]
3. Venkatraman ES, Begg CB. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika. 1996; 83:835–848.
4. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med. 2008; 27:157–172. [PubMed: 17569110]
5. Seshan VE, Gönen M, Begg CB. Comparing ROC curves derived from regression models. Stat Med. 2013; 32:1483–93. [PubMed: 23034816]