5-7-2007

Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting

Holly Janes
Johns Hopkins University, hjanes@jhsph.edu

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

A popular area of medical research today is aimed at the development of markers for classifying subjects as diseased or disease-free, high or low risk, or in terms of treatment response or some other future event. These markers may be the results of, for example, genetic or proteomic evaluations, imaging techniques, bacterial culture, or risk factor information. Often times, there are other factors which affect marker levels. For example, prostate-specific antigen (PSA), a biomarker widely used to screen men for prostate cancer, tends to increase with age. Many markers are also affected by aspects of the test procedure, test setting, or test operator; attributes of the specimen collection or storage method (e.g., storage time); or "center effects" in multi-center studies. While adjustment for covariates is commonplace in therapeutic and etiologic studies, the issue of covariate effects is not well appreciated in the classification setting. In this paper, we demonstrate the need for covariate adjustment and describe statistical methods that can be used to accomplish this. We also distinguish covariate adjustment from several other closely related but fundamentally different concepts, including matching, risk score estimation, and incremental value. Finally, we provide practical recommendations for determining when and how to adjust for covariates, and include links to software that can be used to implement these techniques.

Why Adjust for Covariates?

The classification accuracy of a continuous marker, Y, is its ability to distinguish between two groups defined by an outcome, which we loosely call 'cases' and 'controls'. Classification accuracy is most commonly quantified using the ROC curve, a plot of

the true positive fraction (TPF; sensitivity) versus the false positive fraction (FPF; 1 - specificity) for the set of rules which classify an individual as "test positive" if their marker value is above a threshold, c, for all possible thresholds. The ROC curve quantifies the separation between the case and control marker distributions. It puts markers on a common scale, thus facilitating comparing markers and comparing results across studies. A sample ROC curve is shown in Figure 1.

Confounding occurs in evaluating classification accuracy when there is a covariate which is associated with both the marker and the binary outcome, D. In the presence of such a covariate, the traditional pooled ROC curve, which combines all case observations together and all control observations together regardless of covariate value, is biased. Consider the example shown in Figure 1, scenario 1 (reproduced from (1)). A binary covariate (Z) is associated with both the outcome and the marker. For concreteness, suppose Z is an indicator of study center, where the proportion of cases differs between the two centers. Observe that the pooled ROC curve for Y is overoptimistic relative to the common covariate-specific ROC curve, since the center with the most cases also tends to have higher marker levels. Failing to adjust for the covariate (center) leads to an overoptimistic measure of marker performance.

But bias can occur even when the covariate is associated with the marker, but not the binary outcome. Suppose that in the two-center study above the proportion of cases is the same in the two centers. Observe in Figure 1, scenario 2 that the pooled ROC curve for Y is now attenuated relative to the common covariate-specific ROC curve. In this case, failing to adjust for the covariate leads to an underoptimistic measure of

marker performance. This is directly analogous to studying the association between a predictor and an outcome, in the presence of a covariate that is associated with the predictor but independent of the outcome. The unadjusted odds ratio is attenuated (2, 3). Unbiased estimation of the odds ratio for the predictor requires adjustment for the covariate.

Observe in the scenarios depicted in Figure 1 that, when a covariate affects the distributions of the marker values but not the covariate-specific separation between cases and controls, the separation seen in the pooled data is incorrect. Adjustment for the covariate is necessary in order to appropriately compare the case and control marker distributions. The covariate-adjusted ROC curve, written $\mathcal{A}$ROC, is a measure of covariate-adjusted classification accuracy (1). Conceptually, it is a stratified measure of performance. When the performance of the marker is the same across covariate groups (in other words, the covariate is not an effect modifier), the $\mathcal{A}$ROC is the common covariate-specific ROC curve, which describes the performance of the marker in a population with fixed covariate value. See the solid ROC curve in Figure 1(b). It is analogous to the adjusted odds ratio in an association study. Figure 2(a) shows the age-adjusted ROC curve for PSA, estimated using data from the Physicians' Health Study (4). This ROC curve describes the ability of PSA to discriminate between prostate cancer cases and controls of the same age.

Estimation of the simple ROC curve involves standardizing case marker observations with respect to the control reference distribution, and then calculating the cumulative distribution function of these standardized marker values (5–7). Estimation

of the $\mathcal{A}$ROC is identical except that case observations are standardized with respect to the control distribution with the same covariate value as the case (1, 7). Additional details on estimating the $\mathcal{A}$ROC, including links to software, are included in the appendix.

The Need to Adjust for Covariates When Comparing Markers

In therapeutic studies with paired designs, the effects of patient-specific characteristics are controlled by measuring both predictors on the same subject. For example, in a crossover trial of two drugs, a comparison of the responses under the two drugs does not require adjustment by patient-level covariates because of the paired aspect of the design. However, in evaluating classification accuracy, we are not directly comparing the markers; rather, we are comparing their ROC curves, or the separation between the associated case and control distributions. Therefore, covariate adjustment is still necessary. Consider the example shown in Figure 3, where Y_1 and Y_2 are two markers measured on the same set of subjects. Suppose that Y_1 and Y_2 have the same inherent performance (ROC curve), but Y_1 is affected by a binary covariate, Z, say study site, while Y_2 is not. Observe that the pooled ROC curves incorrectly indicate that Y_2 outperforms Y_1, since the ROC curve for Y_1 is attenuated.

Markers can be compared with regard to covariate-adjusted classification accuracy using any of the commonly used ROC summary indices. For example, the area under

the adjusted ROC curve ($\mathcal{A}$AUC), the partial area under the adjusted ROC curve (p$\mathcal{A}$AUC), estimated sensitivity at a fixed specificity, or estimated specificity at a fixed sensitivity can be used as summary measures. Links to software for estimating and comparing these indices are provided in the appendix.

Why Matching is not Enough

Matching is a design technique which is commonly used when there are covariate effects on classification accuracy. Cases are randomly sampled, and controls are matched to the cases with respect to covariates known to be associated with the marker and the outcome. Such matching is an attempt to control for confounding by these covariates, as illustrated in Figure 1. For example, in the Physicians' Health Study, controls were matched to cases with respect to age in order to eliminate the contribution of age to the apparent discriminatory accuracy of PSA (4). However, matching alone does not solve the problem of confounding.

In etiologic studies, it has long been understood that matching does not eliminate confounding. Odds ratios estimated from a matched study must be adjusted for the matching covariates in the analysis (2, 3). Without adjustment, the odds ratios are biased towards unity. The real role of matching is for efficiency gain in estimating these odds ratios (2, 3).

Directly analogous results have been found in the classification setting (8). That is, matching does not eliminate the confounding. Rather, it converts the confounded pooled ROC curve for the marker into an attenuated ROC curve. Consider the example

shown in Figure 1 scenario 1, where Z (e.g., study center) is associated with both the marker and the outcome. We observe that the unadjusted pooled ROC curve for the marker is overoptimistic. A matched design forces Z to be independent of the outcome in the data (i.e., the same proportion of cases in the two study centers), as in Figure 1 scenario 2. We see that the pooled ROC curve under such a design is still biased, attenuated towards the 45° line. The covariate-adjusted ROC curve correctly estimates the common covariate-specific ROC curve.

Figure 2(a) contrasts the age-adjusted ROC curve for PSA with the unadjusted pooled ROC curve for PSA in the age-matched Physicians' Health Study. Observe that the unadjusted ROC curve in the matched data is generally lower than the age-adjusted ROC curve, as it does not account for the matching.

As in etiologic studies, the real role for matching in studies of classification accuracy is for efficiency gain. Matching has been shown to be a maximally efficient design in many instances (8).

Other Uses for Covariates: Risk Score Estimation and Incremental Value

Covariate adjustment is commonly confused with other uses for covariates in analyses of classification accuracy. We first consider risk score estimation. The risk score is the probability of the outcome (e.g., disease) as a function of marker and covariate information (i.e., $P[D = 1|Y, Z]$). This function is commonly estimated using logistic

regression where the outcome is regressed on one or more markers and other covariate information,

$$\log \text{odds}\, P[D = 1|Y, Z] = \beta_0 + \beta_1 Y + \beta_2 Z.$$

We emphasize that the ROC curve for the risk score is different from the covariate-adjusted ROC curve for the marker. The ROC curve for the risk score describes the ability of the combination of marker and covariates to discriminate between cases and controls. Observe that this combination allows Z to contribute to discrimination. Hence, the risk score may perform well even if Y is a poor classifier, in particular if Z discriminates well. Figure 4 (reproduced from (1)) displays two examples where the ROC curve for the risk score is much higher than the $\mathcal{A}$ROC. In panel (a), Z is a good classifier but Y is not, and the two are relatively uncorrelated. The risk score performs well, but the covariate-adjusted ROC curve for Y, i.e. the ROC curve for Y stratified by Z, is low because it relates to the discriminatory accuracy of Y. In panel (b), both Y and Z are good classifiers which are highly correlated. The risk score performs well, as expected since it should be at least as good as either marker on its own. However, after adjustment for Z the ROC curve for Y is low because within a population where Z is fixed, Y is not a good discriminator. Most of its marginal discrimination is explained by Z, with which it is highly correlated.

Another concept commonly confused with covariate adjustment is incremental value. The incremental value of the marker over the covariates is the amount of discriminatory accuracy that the marker adds to the covariates. It is quantified by comparing the ROC curve for the risk score, the optimal combination of marker and covariates (9), to

the ROC curve for the covariates alone (10). Figure 5 (modified from (8)) shows two examples which demonstrate that the covariate-adjusted performance of a marker is different from its incremental value. In panel (a), the incremental value of the marker is large, but the covariate-adjusted ROC curve is low, and in panel (b), the incremental value is small but the covariate-adjusted ROC curve is high. This interesting finding represents another contrast between studies of association and studies of classification. In association studies, the contribution of one predictor over and above another is its adjusted effect on the outcome. In studies of classification accuracy, these are two different constructs, once again a consequence of the fact that we are not adjusting for the effect of Z on the marker itself, but on the separation between the case and control marker distributions (the ROC curve).

When Covariates Affect Discrimination

Covariates which affect the discriminatory accuracy of the marker (the ROC curve) are analogous to effect modifiers in the association setting. Common examples are severity of disease and expertise of the test operator. With such covariates, a separate ROC curve should be estimated for each covariate group. Covariate adjustment is often a necessary first step in estimating covariate-specific ROC curves, in order to adjust for the effects of the covariate on marker observations among controls. Figure 6 displays a marker, Y, whose accuracy depends on a binary covariate, Z. For concreteness, suppose again that Z is an indicator of study center. Now, however, differences in test procedures between centers affect marker performance (the separation between the case

and control distributions) as well as the marker distributions. Observe that Y is much more accurate when $Z = 0$ than when $Z = 1$. Marker observations among controls also depend on Z (center), necessitating covariate adjustment, or standardization of case marker observations relative to the appropriate covariate-specific control distribution. Panel (b) shows the covariate-specific ROC curves for Y estimated with and without adjustment for Z (center). Observe that the estimates of covariate-specific performance are biased when Z is not adjusted. Case observations in the $Z = 1$ group are much less unusual relative to the general (pooled) control marker distribution than they are relative to the $Z = 1$ controls, leading to an attenuated ROC curve without adjustment for Z. Case observations in the $Z = 0$ group are more unusual relative to the general (pooled) control marker distribution than they are relative to the $Z = 0$ controls, leading to an overoptimistic ROC curve without adjustment for Z.

The covariate-adjusted ROC curve is still useful when the covariate affects discrimination. It turns out that the $\mathcal{A}$ROC is a weighted average of the covariate-specific ROC curves, with weights corresponding to the proportion of cases in each covariate group (1). Observe in Figure 6(c) that the $\mathcal{A}$ROC for the marker lies in between the two covariate-specific ROC curves. It can be interpreted as the average performance of the marker across the two covariate groups. We see then that the $\mathcal{A}$ROC is directly analogous to the Mantel-Haentzel adjusted odds ratio: it is the common covariate-specific ROC curve when Z does not affect discrimination, and a weighted average more generally. It is useful in small studies when covariate-specific ROC curves cannot be estimated with precision, and also provides a single summary of covariate-adjusted

performance for comparing markers.

ROC regression methods can be used to test for whether or not covariates affect discrimination (6). In an ROC regression model, the parameters which describe the effect of Z on the ROC curve can be tested for statistical significance. An ROC regression model was fit for PSA using Physicians' Health Study data; the resulting age-specific ROC curves are shown in Figure 2(b). Observe that there is essentially no variation in discrimination across the age groups. The hypothesis test of the equivalence of the ROC curves is not significant ($p = 0.98$), implying that the $\mathcal{A}$ROC (also shown in Figure 2(b)) is the common age-specific ROC curve for PSA.

Discussion and Practical Recommendations

Rigorous evaluation of new markers being developed for medical classification purposes is essential, and adjustment for covariates is an important component of this evaluation. Adjustment is necessary for covariates that affect marker observations among controls. Failing to adjust for such covariates will lead to biased measures of marker performance. Covariate adjustment is also essential for comparing markers, even under a paired design, as unadjusted comparisons are biased in general. Neither does matching eliminate the need for covariate adjustment; unadjusted measures of marker performance in matched studies are generally attenuated.

The final measure of covariate-adjusted classification accuracy will depend on whether the covariates also affect discriminatory accuracy (are effect modifiers). The $\mathcal{A}$ROC and its associated summary indices are appropriate when the covariates do not af-

fect discrimination, as well as in small studies and when comparing markers. If there is heterogeneity in the accuracy of the marker across covariate groups, estimating covariate-specific ROC curves should be the ultimate goal.

In practice, we suggest exploring associations between all measured covariates and the marker among controls. If any associations are apparent, these covariates should be used for adjustment. Alternatively, all measured covariates can be used for adjustment, and the associated $\mathcal{A}$ROC compared with the pooled ROC to determine whether adjustment makes any difference. In our experience, covariates must be very strongly associated with the marker in order to cause appreciable bias in the pooled ROC curve.

A relatively minor concern is the potential for a loss of efficiency associated with adjusting for covariates which are in fact independent of marker observations among controls, but which appear to be associated in a given data set by random chance. Interestingly, matching with respect to covariates prevents this loss of efficiency; the pooled and covariate-adjusted ROC curves are equally efficient under a matched design (8).

Covariate adjustment is appropriate for covariates whose associations with the marker and the outcome are considered, in some sense, a nuisance. Covariates which are considered markers in their own right should be allowed to contribute to discrimination and should be combined with the marker in the risk score.

Acknowledgements

The authors thank Gary Longton for creating the figures.

Appendix

We have developed Stata programs for estimating, plotting, and comparing covariate-adjusted ROC curves. These programs can be found at the Diagnostics and Biomarkers Statistical Center (DABS) website, http://www.fhcrc.org/science/labs/pepe/dabs.

Estimation of the $\mathcal{A}$ROC involves standardizing case observations with respect to the appropriate covariate-specific control distribution. A standardized case observation is called its "placement value" (5–7, 11, 12). The cumulative distribution (CDF) of the case placement values is the $\mathcal{A}$ROC. Hence, estimation of the $\mathcal{A}$ROC can be divided into two steps: 1) calculate the placement values; 2) estimate their CDF.

In estimating the placement values, one must first decide how to model the covariates. Under existing approaches, this can be done by stratifying on the covariates or by assuming that they act linearly on control marker observations. Next, one must decide how to calculate the corresponding placement values, either empirically or assuming a normal model for control marker observations. Finally, the CDF of the placement values can be calculated empirically or based on the assumption of a binormal ROC curve (13–16).

The $\mathcal{A}$AUC and p$\mathcal{A}$AUC can also be viewed as functions of case placement values (11, 17). We estimate the $\mathcal{A}$AUC as 1 minus the sample mean of the case placement values, and the p$\mathcal{A}$AUC as the sample mean of the "restricted" case placement values. Another interesting summary measure is the estimated TPF at a fixed FPF. Inference about these $\mathcal{A}$ROC summary measures is accomplished using bootstrapping. Clustered data can be accommodated by bootstrapping the clusters.

References

1. Janes H, Pepe M. Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve. Technical Report 283, UW Biostatistics Working Paper Series, 2006. Available at: http://www.bepress.com/uwbiostat/paper283.

2. Breslow N. Handbook of Epidemiology. Springer, 2005.

3. Rothman KJ, Greenland S. Modern Epidemiology. Philadelphia: Lippincott Williams and Wilkins, 1998.

4. Gann PH, Hennekens CH, Stampfer MJ. A prospective evaluation of plasma prostate-specific antigen for detection of prostatic cancer. Journal of the American Medical Association 1995;273:289–94.

5. Pepe M, Longton G. Standardizing diagnostic markers to evaluate and compare their performance. Epidemiology 2005;16:598–603.

6. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press, 2003.

7. Huang Y, Pepe M. Biomarker evaluation using the controls as a reference population. Technical Report 306, UW Biostatistics Working Paper Series, 2007. Available at: http://www.bepress.com/uwbiostat/paper306.

8. Janes H, Pepe MS. Matching in studies of classification accuracy: Implications for bias, efficiency, and assessment of incremental value. Biometrics ;(in press).

9. McIntosh M, Pepe M. Combining several screening tests: Optimality of the risk score. Biometrics 2002;58:657–64.

10. Pepe M, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology 2004;159:882–90.

11. Hanley JA, Hajian-Tilaki KO. Sampling variability of non-parametric estimates of the areas under receiver operating characteristic curves: An update. Academic Radiology 1997;4:49–58.

12. Pepe MS, Cai T. The analysis of placement values for evaluating discriminatory measures. Biometrics 2004;60:528–35.

13. Swets JA. Indices of discrimination or diagnostic accuracy: Their ROCs and implied methods. Psychological Bulletin 1986;99:100–17.

14. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods From Signal Detection Theory. Academic Press, 1982.

15. Hanley JA. The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. Statistics in Medicine 1996;15:1575–85.

16. Hanley JA. The robustness of the 'binormal' assumptions used in fitting ROC curves. Medical Decision Making 1988;8:197–203.

17. Dodd L, Pepe MS. Partial AUC estimation and regression. Biometrics 2003; 59:614–23.

Figure 1: A simulated marker Y and binary covariate $Z = 0, 1$. Under scenario 1, Z is associated with the outcome: $P[Z = 1|D = 0] = 0.10$ and $P[Z = 1|D = 1] = 0.50$. Under scenario 2, Z is independent of the outcome: $P[Z = 1|D = 0] = P[Z = 1|D = 1] = 0.50$. (a) The densities of Y conditional on $Z = 0$, conditional on $Z = 1$, in the pooled data under scenario 1, and in the pooled data under scenario 2. A common threshold of 2.5 is indicated. (b) The common covariate-specific ROC curve, the pooled ROC curve under scenario 1, and the pooled ROC curve under scenario 2. The performances of the thresholding rule are indicated.

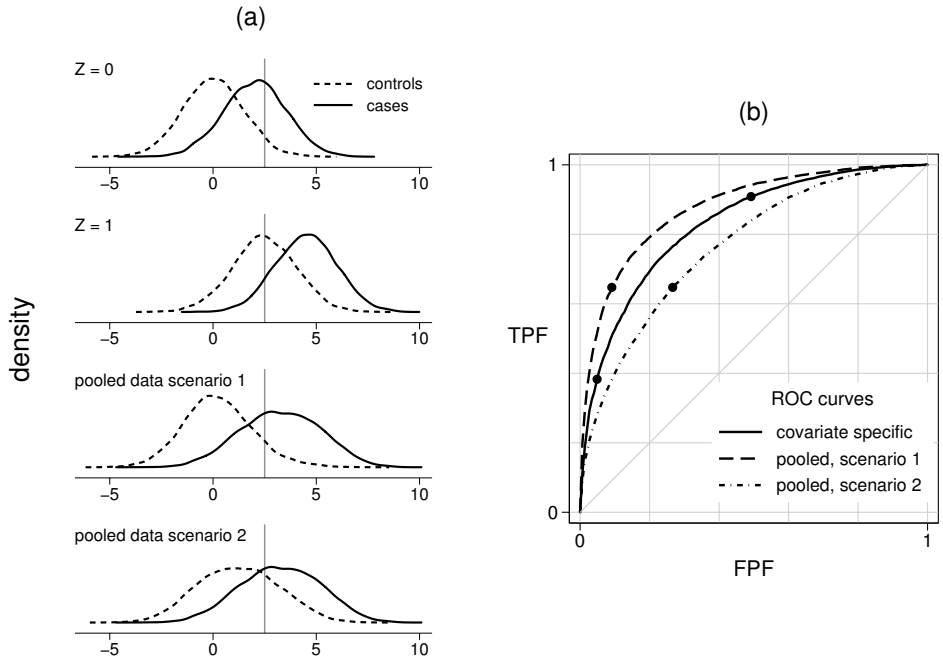

Figure 2: ROC curves for PSA in the PHS data. (a) The age-adjusted ROC curve and the pooled unadjusted ROC curve in the matched data; (b) Age-specific and age-adjusted ROC curves.

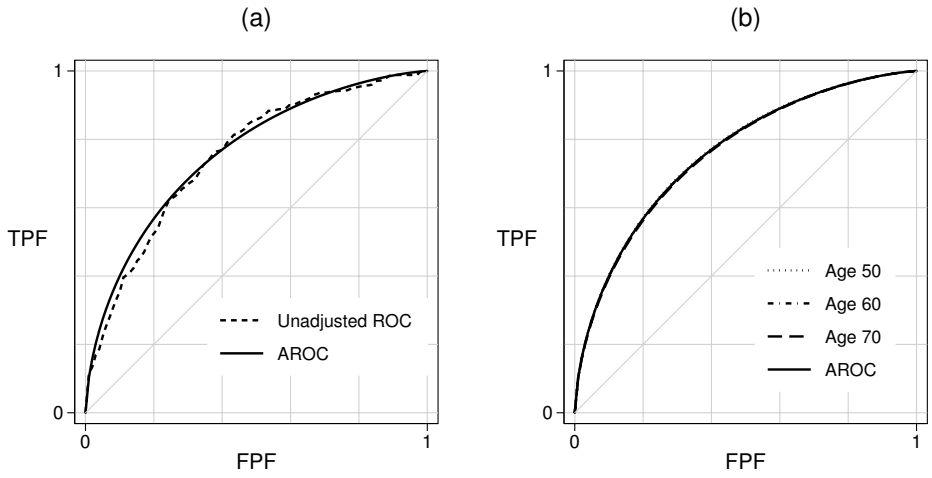

Figure 3: Two simulated markers Y_1 and Y_2 which have the same performance. Y_1 depends on a binary covariate $Z = 0, 1$, while Y_2 does not. The binary outcome does not depend on the covariate. (a) The densities of Y_1 conditional on $Z = 0$, conditional on $Z = 1$, and in the pooled data. A common threshold of 2.5 is indicated. (b) The densities of Y_2 conditional on $Z = 0$, conditional on $Z = 1$, and in the pooled data. (c) The common covariate-specific ROC curve for Y_1 and Y_2, the pooled ROC curve for Y_1, and the pooled ROC curve for Y_2. The performances of the thresholding rule are indicated.

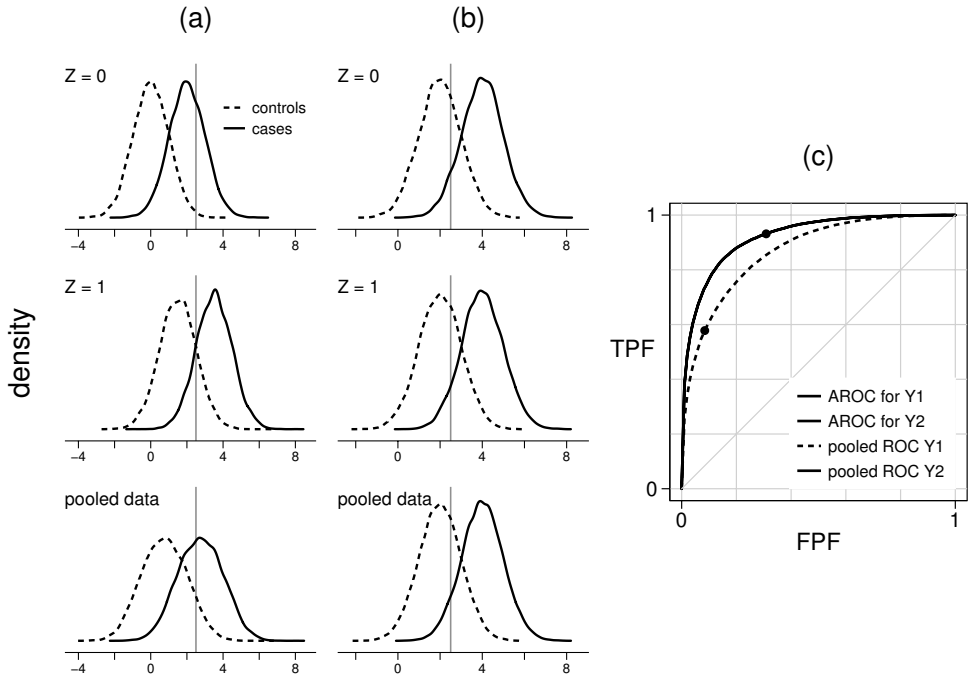

Figure 4: Two simulated examples to illustrate that the ROC curve for the risk score, $R = P[D = 1|Y, Z]$, is different from the common covariate-specific ROC curve. The ROC curve for R and the common covariate-specific ROC curve are shown. (a) Z is a good classifier but Y is not, and the two are relatively uncorrelated. (b) Both Y and Z are good classifiers, and are highly correlated. Reproduced from (1).

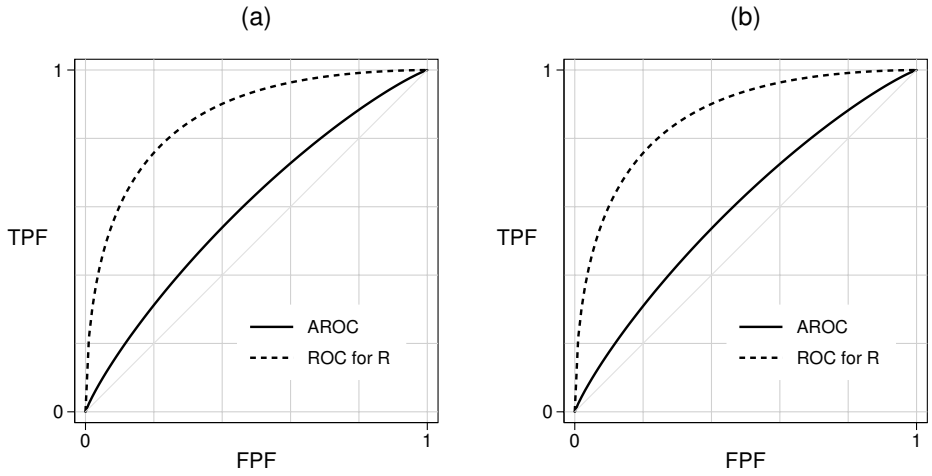

Figure 5: Two simulated examples, illustrating that the covariate-adjusted ROC curve is not related to the incremental value. In each example, the covariate-adjusted ROC curve is shown, along with the ROC curves for the risk score, $R = P[D = 1|Y, Z]$, and for Z alone. (a) The incremental value is large, and the covariate-adjusted ROC curve is low. (b) The incremental value is small, and the covariate-adjusted ROC curve is high. Modified from (8).

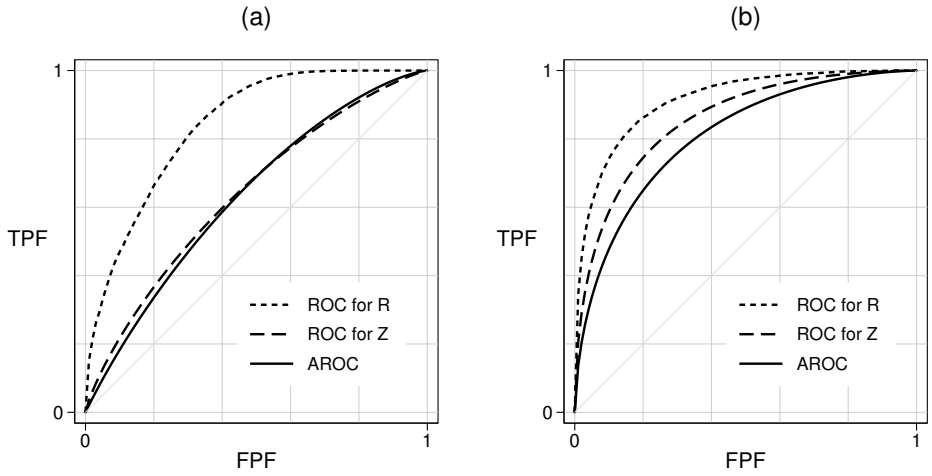

Figure 6: A simulated marker Y and binary covariate $Z = 0, 1$. The performance of Y depends on Z, but Z is independent of the outcome. (a) The densities of Y conditional on $Z = 0$, conditional on $Z = 1$, and in the pooled data. A common threshold of 2.5 is indicated. (b) The covariate-specific ROC curves for Y with and without adjustment for Z. The performances of the threshold are indicated. (c) The covariate-specific and covariate-adjusted ROCs curves for Y.