The Stata Journal (2009)
9, Number 1, pp. 17–39

Accommodating covariates in receiver operating characteristic analysis

Holly Janes
Fred Hutchinson Cancer Research Center
Seattle, WA
hjanes@fhcrc.org

Gary Longton
Fred Hutchinson Cancer Research Center
Seattle, WA
glongton@fhcrc.org

Margaret S. Pepe
Fred Hutchinson Cancer Research Center
Seattle, WA
mspepe@u.washington.edu

Abstract. Classification accuracy is the ability of a marker or diagnostic test to discriminate between two groups of individuals, cases and controls, and is commonly summarized by using the receiver operating characteristic (ROC) curve. In studies of classification accuracy, there are often covariates that should be incorporated into the ROC analysis. We describe three ways of using covariate information. For factors that affect marker observations among controls, we present a method for covariate adjustment. For factors that affect discrimination (i.e., the ROC curve), we describe methods for modeling the ROC curve as a function of covariates. Finally, for factors that contribute to discrimination, we propose combining the marker and covariate information, and we ask how much discriminatory accuracy improves (in incremental value) with the addition of the marker to the covariates. These methods follow naturally when representing the ROC curve as a summary of the distribution of case marker observations, standardized with respect to the control distribution.

Keywords: st0155, roccurve, comproc, rocreg, receiver operating characteristic analysis, ROC, covariates, sensitivity, specificity

1 Introduction

The classification accuracy of a marker, Y, is most commonly described by the receiver operating characteristic (ROC) curve, which is a plot of the true positive rate (TPR) versus the false positive rate (FPR) for the set of rules that classify an individual as "test-positive" if $Y \geq c$, where the threshold, c, is varied over all possible values (Pepe et al. 2001; Baker 2003). Equivalently, the ROC curve can be represented as the cumulative distribution function (c.d.f.) of the case marker observations, standardized with respect to the control distribution (Pepe and Cai 2004; Pepe and Longton 2005). The standardized marker observations, or percentile values, are written as $\text{PV}_D = F(Y_D)$, where F is the left-continuous c.d.f. of Y among controls, and Y_D denotes a case marker observation. The ROC curve at an FPR of f is

$$\text{ROC}(f) = P(1 - \text{PV}_D \leq f)$$

In many settings, covariates should be incorporated into the ROC analysis. There are covariates that impact the marker distribution among controls. For example, "center effects" in multicenter studies may affect marker observations. In section 2, we describe methods for adjusting the ROC curve for such covariates. The associated Stata commands are `roccurve` and `comproc` (Pepe, Longton, and Janes 2009). Other covariates may affect the inherent discriminatory accuracy of the marker (i.e., the ROC curve). For example, disease severity often impacts marker accuracy, with less severe cases being more difficult to distinguish from controls. In section 3, we describe an ROC regression method that allows the ROC curve to depend on covariates. The associated Stata command is `rocreg`, which we introduce in section 3.3. Finally, there are covariates that contribute to discrimination. For example, baseline risk factors for disease provide some ability to discriminate between cases and controls. A common question is how much discriminatory accuracy (i.e., incremental value) the marker adds to the known classifiers. In section 4, we describe methods for evaluating incremental value.

This article is a companion to another article in this issue (Pepe, Longton, and Janes 2009); the companion article describes the use of `roccurve` and `comproc` for estimating and comparing ROC curves without incorporating covariate information.

2 The covariate-adjusted ROC curve

2.1 Motivation and concept

Consider a covariate, Z, that affects the distribution of the marker among controls. Figure 1 shows hypothetical data for a continuous marker, Y; a binary outcome, D; and a binary covariate, Z. The data can be downloaded from the Diagnostic and Biomarkers Statistical (DABS) Center web site (http://labs.fhcrc.org/pepe/dabs/). Suppose for concreteness that Z is an indicator of study center. Marker observations among controls ($D = 0$) tend to be higher in center 1 compared with center 0, but the inherent accuracy of the marker (the ROC curve) is the same in the two centers. Consider the pooled ROC curve for Y; this curve combines all case observations together and all control observations together, regardless of study center. In figure 1, observe that when the proportion of cases varies across centers (scenario 1), the pooled ROC curve for Y is overly optimistic compared with the ROC curve for Y in each center. Even when Z is independent of the outcome (i.e., even when the proportion of cases is held constant across centers, as in scenario 2), the pooled ROC curve is biased; this time, it is attenuated with respect to the center-specific ROC curve. This suggests that covariates that impact marker observations among controls should be statistically adjusted in the ROC analysis.

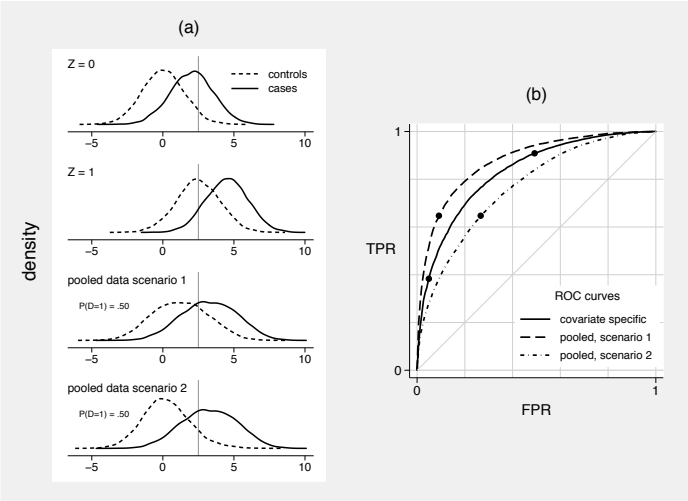

Figure 1. A simulated marker, Y, and binary covariate, $Z = 0, 1$. Under scenario 1, Z is associated with the outcome: $P(D = 1 \mid Z = 0) = 0.36$ and $P(D = 1 \mid Z = 1) = 0.83$. Under scenario 2, Z is independent of the outcome: $P(D = 1 \mid Z = 0) = P(D = 1 \mid Z = 1) = 0.50$. (a) Shown are the densities of Y conditional on $Z = 0$, then conditional on $Z = 1$, then in the pooled data under scenario 1, and finally in the pooled data under scenario 2. A common threshold is indicated. (b) Shown are 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 common threshold rule are indicated.

We propose a covariate-adjusted measure of classification accuracy called the covariate-adjusted ROC curve, or the $\mathcal{A}$ROC (Janes and Pepe Forthcoming, 2008). Conceptually, this is a stratified measure of marker performance. It is defined as

$$\mathcal{A}\text{ROC}(f) = P(1 - \text{PV}_{DZ} \leq f)$$

where PV stands for percentile value, and $\text{PV}_{DZ} = F_Z(Y_{DZ})$ represents the case observation with the covariate value Z (Y_{DZ}) standardized with respect to the control population with the same value of Z. When the performance of the marker is the same across populations with different values of Z, as in figure 1, the $\mathcal{A}$ROC is the common covariate-specific ROC curve. More generally, it is a weighted average of covariate-specific ROC curves (Janes and Pepe Forthcoming). Equivalently, the $\mathcal{A}$ROC is the ROC curve for Y when Z-specific thresholds are used for classification. The thresholds, c_Z, are chosen to ensure that the covariate-specific FPR, $\text{FPR}_Z(c_Z)$, is common across levels of Z.

2.2 Estimating the $\mathcal{A}$ROC

Estimation of the $\mathcal{A}$ROC proceeds in two steps:

1. Estimate F_Z, the distribution of the marker in controls as a function of Z. For each case subject, i, calculate the PV: $\text{PV}_{DZ_i} = F_{Z_i}(Y_{DZ_i})$.
2. Estimate the c.d.f. of the case PVs.

Estimating F_Z begins with specifying how Z acts on the distribution of Y among controls. For example, a linear model could be specified:

$$Y = \beta_0 + \beta_1 Z + \epsilon$$

The random error, ϵ, could be assumed to be normally distributed, $\epsilon \sim N(0, \sigma^2)$, which would lead to the case PVs

$$\widehat{\text{PV}}_{DZ} = \Phi\{(Y - \widehat{\beta}_0 - \widehat{\beta}_1 Z)/\widehat{\sigma}\}$$

where Φ is the standard normal c.d.f., and $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\sigma}$ are estimates from the linear model. Alternatively, the error distribution could be estimated empirically by using the residuals from the linear model as in Heagerty and Pepe (1999). This would lead to the PVs

$$\widehat{\text{PV}}_{DZ} = \widehat{F}\{(Y - \widehat{\beta}_0 - \widehat{\beta}_1 Z)/\widehat{\sigma}\}$$

In addition to allowing Z to act linearly on marker observations among controls, the `roccurve` command allows for stratifying on Z. Here again the distribution of Y among controls conditional on Z can be estimated empirically or by assuming a normal distribution.

Once the PVs have been calculated, their c.d.f. must be estimated. This estimation step is described in more detail in the companion article (Pepe, Longton, and Janes 2009). Briefly, the c.d.f. can be estimated empirically, or a parametric distribution can be assumed. The `roccurve` command allows the parametric forms

$$\text{ROC}(f) = P(1 - \text{PV}_{DZ} \leq f) = g\{\alpha_0 + \alpha_1 g^{-1}(f)\}$$

where $g = \Phi$ is the standard normal c.d.f., or $g(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ is the logistic function. These forms give rise to binormal (Dorfman and Alf 1969) and bilogistic (Ogilvie and Creelman 1968) ROC curves.

To fit the ROC model, a discrete set of FPR points, $f_1, \ldots, f_{n_p}$, is chosen. These points can span the interval $(0, 1)$ or a subinterval of interest, (a, b). For each case observation, a set of n_p records is created. The kth record for the ith subject includes the binary outcome $U_{ki} = \mathbf{I}_{(1-\widehat{\text{PV}}_{DZ_i} \leq f_k)}$ and the covariate $g^{-1}(f_k)$. A binary regression model with the link function g, the outcome U, and the covariate $g^{-1}(f)$ provides estimates of (α_0, α_1) (Alonzo and Pepe 2002).

We bootstrap the data to obtain standard errors for the estimated $\mathcal{A}_{\text{ROC}}$. The data should be resampled according to the design of the study; for a case–control study, this means resampling separately within case and control strata. If the data are clustered, the clusters should be the resampling units.

Consider as an example data from a neonatal audiology study designed to evaluate the accuracy with which three audiology tests identify hearing impairment in newborns (Norton, Wang, and Ai 2004). The data can be downloaded from the DABS Center web site, or it can be loaded directly into Stata by typing

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2
```

Test results for hearing-unimpaired ears may depend on the age and gender of the child. Figure 2 (on page 23) shows the estimated age- and gender-adjusted ROC curves for the marker DPOAE. Several estimation options are shown. The first estimator assumes a linear model for marker measurements among controls,

$$Y = \beta_0 + \beta_1 Z_{\text{age}} + \beta_2 Z_{\text{gender}} + \epsilon$$

where the error distribution is estimated empirically. The c.d.f. of the estimated PVs,

$$\widehat{\text{PV}}_{DZ_i} = \widehat{F}\{(Y - \widehat{\beta}_0 - \widehat{\beta}_1 Z_{\text{age}_i} - \widehat{\beta}_2 Z_{\text{gender}_i})/\widehat{\sigma}\}$$

is estimated empirically. The second estimator adds the assumption that ϵ is normally distributed, and the third estimator additionally assumes that the ROC curve is binormal. Clustered bootstrapping is used for inference to account for correlation among observations (ears) for the same individual. The ROC curve is somewhat sensitive to the normality assumption at the high end of the marker distribution. Next we describe how to estimate these curves by using the `roccurve` command.

2.3 The roccurve command

Syntax

The syntax for the `roccurve` command is

`roccurve` *disease_var test_varlist* [*if*] [*in*] [`,` *options*]

where *disease_var* is the name of the binary outcome ($D = 1$ for a case and $D = 0$ for a control), and *test_varlist* is the list of markers.

Options

See the companion paper (Pepe, Longton, and Janes 2009) in this issue for details of the options for `roccurve`. Here we focus on the options that relate to covariate adjustment.

Marker standardization. The covariates to be used for adjustment are specified by using the `adjcov(`*varlist*`)` option. The `adjmodel(`*model*`)` option specifies how the covariates are to be used for adjustment; the default is `stratified`, where the control marker distribution is stratified on the covariates. *model* can also be `linear`; here the covariates are assumed to act linearly on the control marker distribution.

Standardized marker values are calculated according to the specification in the `pvcmeth(`*method*`)` option. *method* can be `empirical` (the default), where the control marker distribution is estimated empirically conditional on the covariates, or `normal`, where the control marker is assumed to have a normal distribution conditional on the covariates.

ROC calculation. `rocmeth(`*method*`)` specifies whether `nonparametric` (empirical ROC, the default) or `parametric` ROCs are to be calculated. The `link(`*function*`)` option is required for a parametric ROC model; a binormal model is fit with `link(probit)`, and a bilogistic model is fit with `link(logistic)`. For a parametric ROC model, the `interval(`a b n_p`)` option can be used to specify that the model is to be fit at n_p points over the restricted FPR interval (a, b).

Sampling variability. Bootstrapping is used for inference. By default, the data are resampled conditional on the binary outcome. The `noccsamp` option specifies that data be resampled without regard to the outcome. The `nostsamp` option specifies that sampling be done without regard to covariate strata; by default, when covariates are used for stratification, bootstrap samples are drawn from within covariate strata. The `cluster(`*varlist*`)` option can be used to bootstrap clustered data.

Example

The following code produces the estimators shown in figure 2 (your graphs will look slightly different because we do not show all the options here):

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2
. roccurve d y1, adjcov(currage gender) adjmodel(linear) cluster(id) noccsamp
. roccurve d y1, adjcov(currage gender) adjmodel(linear) pvcmeth(normal)
> cluster(id) noccsamp
. roccurve d y1, adjcov(currage gender) adjmodel(linear) pvcmeth(normal)
> rocmeth(parametric) cluster(id) noccsamp
```

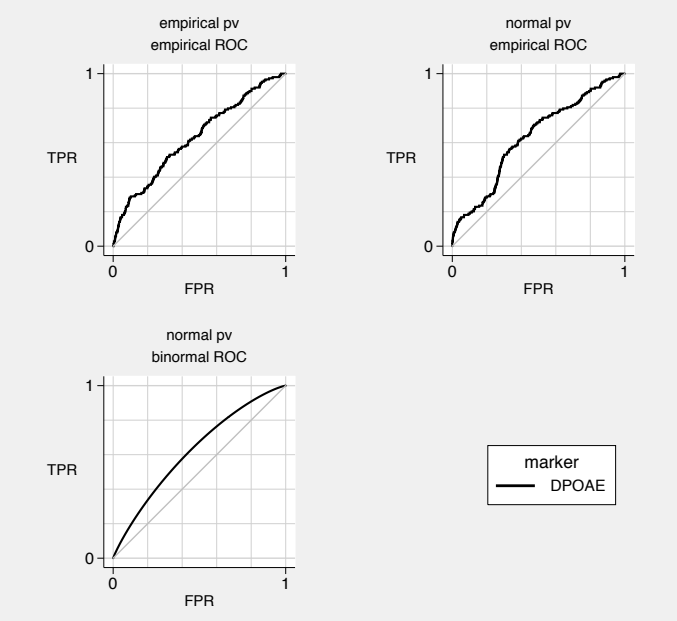

Figure 2. Three different estimates of the age- and gender-adjusted ROC curve for the marker DPOAE based on the Norton, Wang, and Ai (2004) audiology data

ROC summary indices

Summary measures of the ROC curve serve as metrics for comparing markers. The area under the $\mathcal{A}$ROC ($\mathcal{A}$AUC), $\mathcal{A}\text{AUC} = \int_0^1 \mathcal{A}\text{ROC}(f)\,df$, can be interpreted as the probability that, for a random case and control marker observation with the same covariate value, the case observation is higher than the control. This is not a clinically relevant summary of marker performance because the task is not to determine which of a pair of subjects is the case. Moreover, the $\mathcal{A}$AUC summarizes the entire ROC curve when, frequently, only a portion (e.g., only low FPRs) is of interest.

A more clinically meaningful summary measure of the $\mathcal{A}$ROC is the $\mathcal{A}$ROC (TPR) at a fixed FPR $= f$ of interest. This can be interpreted as the percentage of cases detected when the covariate-specific FPRs are held at f. Alternatively, the FPR corresponding to a specific TPR $= \mathcal{A}\text{ROC}^{-1}(t)$ could be reported. This is the common covariate-specific FPR associated with a proportion, t, of cases detected.

The partial area under the $\mathcal{A}$ROC (p$\mathcal{A}$AUC), $\text{p}\mathcal{A}\text{AUC}(f_0) = \int_0^{f_0} \mathcal{A}\text{ROC}(f)\,df$, can be viewed as a compromise between the $\mathcal{A}$AUC and the $\mathcal{A}$ROC at a specified point. The p$\mathcal{A}$AUC has the advantage of focusing on a portion of the $\mathcal{A}$ROC, but it lacks a clinically relevant interpretation.

The $\mathcal{A}$ROC summary measures are estimated in the same way as their counterparts for the traditional ROC curve. The $\mathcal{A}$AUC estimate is the sample average of the case standardized marker values,

$$\widehat{\mathcal{A}\text{AUC}} = \sum_{i=1}^{n_D} \widehat{\text{PV}}_{DZ_i}/n_D \tag{1}$$

where the sum is over the n_D case observations. When the case PVs are estimated nonparametrically (i.e., with stratification on Z), this is a weighted average of the empirical areas under the ROC curves (AUCs) in each covariate stratum. The estimated p$\mathcal{A}$AUC is also an average of standardized marker values (Dodd and Pepe 2003),

$$\text{p}\widehat{\mathcal{A}\text{AUC}}(f_0) = \sum_{i=1}^{n_D} \max\{\widehat{\text{PV}}_{DZ_i} - (1 - f_0), 0\}/n_D \tag{2}$$

When the control marker distribution is estimated empirically, corrections are made for ties between case and control marker observations, as discussed in the companion article (Pepe, Longton, and Janes 2009).

Estimates of $\mathcal{A}$AUC and p$\mathcal{A}$AUC values for parametric ROC models generally require numerical integration and are not produced by our programs. Instead, the parameters are estimated by using empirical averages of PVs, as in (1) and (2). Similarly, we estimate the $\mathcal{A}$ROC at a fixed FPR $= f$ by calculating the proportion of PVs that are greater than $1 - f$ rather than the value estimated by a parametric ROC model.

2.4 Comparing covariate-adjusted ROC curves

Comparisons between $\mathcal{A}$ROCs can be made by using any of the summary indices discussed above. A confidence interval for the difference in summary measures is calculated by using the bootstrap method. A Wald statistic obtained by dividing the observed difference by its standard error is compared to the standard normal distribution to obtain a p-value. Standard errors are obtained by bootstrapping. The `comproc` command is used to compare $\mathcal{A}$ROCs.

2.5 The comproc command

Syntax

The syntax of the `comproc` command is

`comproc` *disease_var test_var1* [*test_var2*] [*if*] [*in*] [`,` *options*]

where *disease_var* is the binary outcome, and *test_var1* and *test_var2* are the two markers to be compared. If only one marker is specified, the requested summary statistics are returned but no comparisons are made.

Options

Marker standardization and bootstrap options are the same as with `roccurve`. The choices of summary measures are `auc`, the $\mathcal{A}$AUC; `pauc(`*f*`)`, the p$\mathcal{A}$AUC; `roc(`*f*`)`, the TPR corresponding to an FPR of f; and `rocinv(`*t*`)`, the FPR corresponding to a TPR of t. The `tiecorr` option can be used to correct for ties between case and control marker observations; it is used by default if `pauc(`*f*`)` is among the summary measures requested.

Example

Consider again the audiology data. Figure 3 shows the ROC curves for the markers DPOAE and TEOAE, both adjusted for age and gender. The covariates are assumed to act linearly on control marker observations, and the marker distributions and ROC curves are estimated empirically. The `comproc` command yields estimates of the associated ROC curves at an FPR of $f = 0.20$, as well as the p$\mathcal{A}$AUC(0.20) and the $\mathcal{A}$AUC, as shown in the output below. We conclude that there is no evidence of a difference in the percentage of cases detected when the FPR is 20%. Comparisons based on the p$\mathcal{A}$AUC(0.20) and the $\mathcal{A}$AUC similarly suggest that there is no difference in performance between the two markers.

(*Continued on next page*)

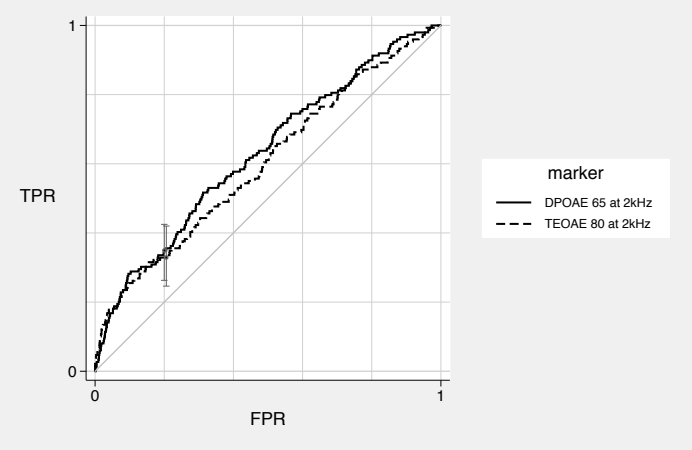

Figure 3. Age- and gender-adjusted ROC curves for the markers DPOAE and TEOAE based on the Norton, Wang, and Ai (2004) audiology data

The `comproc` command applied to the audiology data yields the following results:

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2, clear
(Norton - neonatal audiology data)
. set seed 49049
. comproc d y1 y2, roc(0.2) pauc(0.2) auc adjcov(currage gender) adjmodel(linear)
> cluster(id) noccsamp
        Comparison of test measures
                                test 1: DPOAE 65 at 2kHz
                                test 2: TEOAE 80 at 2kHz
    percentile value calculation method: empirical
        percentile value tie correction: yes
               Covariate adjustment
                                method: linear model
                            covariates: currage
                                        Gender
************
 covariate adjustment - linear model, controls only
 model results for marker: DPOAE 65 at 2kHz
```

Source	SS	df	MS		
Model	2418.56541	2	1209.2827	Number of obs	= 4907
Residual	294662.363	4904	60.0861263	F(2, 4904)	= 20.13
				Prob > F	= 0.0000
				R-squared	= 0.0081
				Adj R-squared	= 0.0077
Total	297080.929	4906	60.5546125	Root MSE	= 7.7515

y1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
currage	-.2032456	.0323905	-6.27	0.000	-.2667455	-.1397458
gender	.2471744	.2229119	1.11	0.268	-.1898327	.6841815
_cons	-1.486659	1.288611	-1.15	0.249	-4.012913	1.039596

```
************
 covariate adjustment - linear model, controls only
 model results for marker: TEOAE 80 at 2kHz
```

Source	SS	df	MS		
Model	2186.03352	2	1093.01676	Prob > F	= 0.0000
Residual	239493.534	4904	48.836365	R-squared	= 0.0090
				Adj R-squared	= 0.0086
Total	241679.567	4906	49.2620398	Root MSE	= 6.9883

Number of obs	=	4907			
F(2, 4904)	=	22.38			

y2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
currage	-.1694143	.0292013	-5.80	0.000	-.2266619	-.1121667
gender	.7014169	.2009638	3.49	0.000	.3074379	1.095396
_cons	-6.348757	1.161733	-5.46	0.000	-8.626274	-4.07124

```
************
  bootstrap samples drawn
   w/o respect to case/control status
  # bootstrap samples: 1000
***************
  AUC estimates and difference,
    test 2 - test 1 (aucdelta)
```

Bootstrap results — Number of obs = 5056; Replications = 1000

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
auc1	.62941998	.0001283	.0258707	.5787143	.6801256	(N)
				.5794999	.6800154	(P)
				.5791458	.6799226	(BC)
auc2	.60102814	.0004614	.02607367	.5499247	.6521316	(N)
				.5511089	.6488042	(P)
				.5509614	.6485289	(BC)
aucdelta	-.02839184	.0003331	.02080347	-.0691659	.0123822	(N)
				-.0676753	.0110209	(P)
				-.067653	.0111732	(BC)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
test of Ho: auc1 = auc2
  z =    -1.4    p =     .17
```

(*Continued on next page*)

```
***************
 pAUC estimates and difference,
   test 2 - test 1 (paucdelta)
  partial AUC for f < .2
Bootstrap results                                 Number of obs      =      5056
                                                  Replications       =      1000
```

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
pauc1	.04624855	.0003327	.00622013	.0340573	.0584398	(N)
				.0345031	.0584833	(P)
				.0343457	.0581308	(BC)
pauc2	.04657379	.0003689	.00671733	.0334081	.0597395	(N)
				.0344294	.0602263	(P)
				.0340786	.0600065	(BC)
paucdelta	.00032524	.0000361	.00458657	-.0086643	.0093148	(N)
				-.0085509	.0096089	(P)
				-.0084957	.0096477	(BC)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
test of Ho: pauc1 = pauc2
  z =    .071    p =     .94
***************
 ROC estimates and difference,
    test 2 - test 1 (rocdelta)
  ROC(f) @ f = .2
Bootstrap results                                 Number of obs      =      5056
                                                  Replications       =      1000
```

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
roc1	.3489933	-.0026902	.04117657	.2682887	.4296979	(N)
				.2658619	.4260527	(P)
				.2681159	.4266667	(BC)
roc2	.32885906	.004282	.04411025	.2424046	.4153136	(N)
				.2507862	.418124	(P)
				.245283	.4113475	(BC)
rocdelta	-.02013424	.0069722	.03972608	-.0979959	.0577274	(N)
				-.0931783	.06618	(P)
				-.1048951	.0522388	(BC)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
test of Ho: roc1 = roc2
  z =    -.51    p =     .61
***************
```

3 ROC regression

3.1 Motivation and concept

Covariates such as disease severity and specimen storage time can do more than impact marker observations among controls. They often also impact the inherent discriminatory accuracy of the marker (i.e., the ROC curve). That is, they affect the separation between the case and the control marker distributions. A hypothetical example is shown in figure 4. The data used can be downloaded from the DABS Center web site. The separation between the case and the control marker distributions is much greater when $Z = 0$ than when $Z = 1$. The covariate also affects the distribution of the marker among controls, necessitating covariate adjustment.

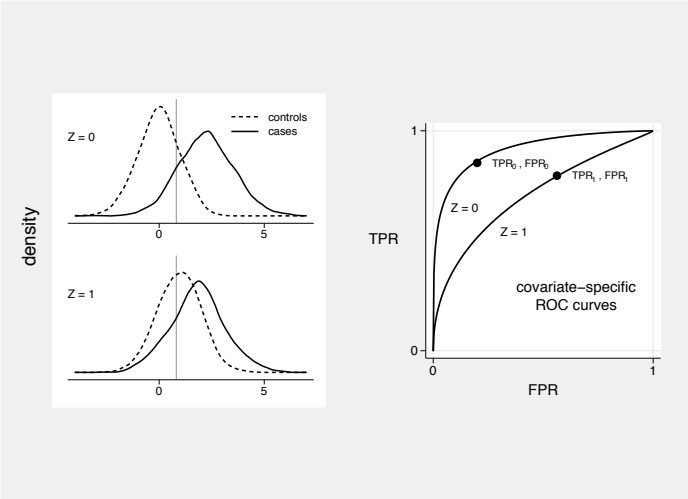

Figure 4. A simulated marker, Y, and binary covariate, $Z = 0, 1$. The marker is more accurate when $Z = 0$ than when $Z = 1$, and marker observations among controls also depend on Z. The performances of a common threshold are indicated.

ROC regression is a methodology that models the marker's ROC curve as a function of covariates (Pepe 2000; Alonzo and Pepe 2002). Implementation proceeds in two steps:

1. Model the distribution of the marker among controls as a function of covariates, and calculate the case PVs.

2. Model their c.d.f. (i.e., the ROC curve) as a function of covariates.

The result is an estimate of the ROC curve for the marker as a function of covariates, or a covariate-specific ROC curve. We emphasize that the covariates used in step 1 for adjustment are those that affect the marker distribution in the control population; these

may or may not differ from the covariates used in step 2 that impact the separation between cases and controls.

3.2 Estimation

Estimation of the control marker distribution as a function of covariates and calculation of the case PVs proceeds in exactly the same manner as with the covariate-adjustment method. The standardization options allowed by `rocreg`, introduced in the next section, are the same as with `roccurve` and `comproc`. The covariates can be assumed to act linearly on marker observations, or stratification can be employed if they are discrete. The PVs can be calculated by empirically estimating the control marker distribution conditional on covariates or by assuming a normal model.

Next a parametric model is specified for the ROC curve (i.e., the c.d.f. of the case PVs) as a function of covariates. The forms allowed by the `rocreg` command are

$$\text{ROC}_Z(f) = P(1 - \text{PV}_{DZ} \leq f) = g\{\alpha_0 + \alpha_1 g^{-1}(f) + \alpha_2 Z + \alpha_3 Z \times g^{-1}(f)\}$$

where $g(\cdot)$ is the standard normal c.d.f. or the logistic function. The parameter α_2 allows the covariates to impact the "intercept" of the ROC curve, while α_3 allows Z to impact the "slope" of the ROC curve. If $\alpha_3 \neq 0$, the covariates have a different impact on the ROC curve at different FPRs. This ROC model gives rise to binormal (Dorfman and Alf 1969) or bilogistic (Ogilvie and Creelman 1968) ROC curves at each fixed value of Z.

To fit the ROC regression model, a discrete set of FPR points, $f_1, \ldots, f_{n_p}$, is chosen. These points can span $(0, 1)$ or a subinterval of interest, (a, b). For each case observation, a set of n_p records is created. The kth record for the ith subject includes the binary outcome $U_{ki} = \mathbf{I}_{(1-\widehat{\text{PV}}_{DZ_i} \leq f_k)}$ and the covariates $g^{-1}(f_k)$, Z_i, and $Z_i \times g^{-1}(f_k)$. A binary regression model with the link function g; the outcome U; and the covariates $g^{-1}(f)$, Z, and $Z \times g^{-1}(f)$ provides estimates of $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ (Alonzo and Pepe 2002). Bootstrapping is used for inference, where the data are resampled according to the design.

For illustration, an ROC regression model was fit for the marker DPOAE by using the audiology data. DPOAE observations among controls are assumed to depend linearly on age and gender, and their distribution is estimated empirically. Age-specific ROC curves are modeled parametrically by using

$$\text{ROC}_Z(f) = \Phi\{\alpha_0 + \alpha_1 \Phi^{-1}(f) + \alpha_2 Z_{\text{age}}\} \tag{3}$$

Estimates of the age-specific ROC curves are calculated by using the parameter estimates $(\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2)$. Figure 5 shows estimated binormal ROC curves for children at 30, 40, and 50 months of age. This figure suggests that the marker is more accurate among older children, but the effect is not statistically significant, as we will see.

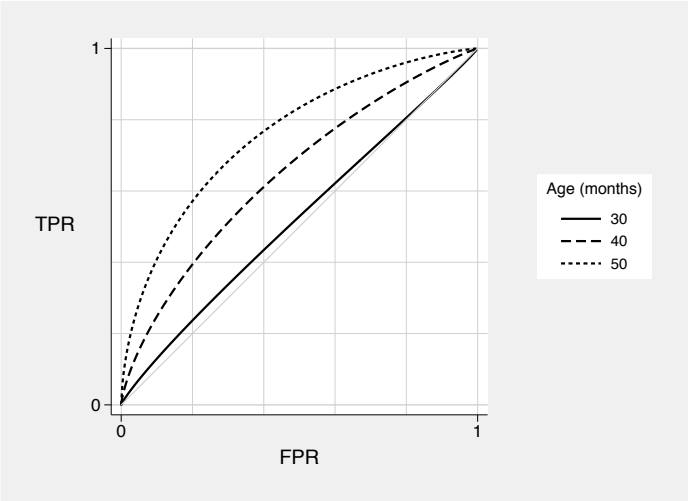

Figure 5. Age-specific ROC curves for the marker DPOAE based on the Norton, Wang, and Ai (2004) audiology data. The ROC curves are adjusted for age and gender.

3.3 The rocreg command

Syntax

The syntax of the `rocreg` command is

`rocreg` *disease_var test_varlist* [*if*] [*in*] [`,` `regcov(`*varlist*`)` `sregcov(`*varlist*`)` `link(`*function*`)` `interval(`*a b* n_p`)` `pvcmeth(`*method*`)` `tiecorr` `adjcov(`*varlist*`)` `adjmodel(`*model*`)` `nsamp(`#`)` `nobstrap` `noccsamp` `nostsamp` `cluster(`*varlist*`)` `resfile(`*filename*`)` `replace` `level(`#`)`]

where *disease_var* is the binary outcome, and *test_varlist* is the list of markers.

Options

ROC regression

`regcov(`*varlist*`)` specifies the variables to be included in the ROC regression model that affect the intercept of the ROC curve.

`sregcov(`*varlist*`)` specifies the variables to be included in the ROC regression model that affect the slope of the ROC curve.

`link(`*function*`)` specifies the ROC generalized linear model (ROC-GLM) link function. *function* can be one of the following:

`probit`, the default, corresponds to the binormal ROC model. That is, $\Phi^{-1}\{\text{ROC}(f)\} = \text{intercept} + \text{slope} \times \Phi^{-1}(f)$, where Φ is the standard normal c.d.f.

`logit` corresponds to the bilogistic ROC model. That is, $\text{logit}\{\text{ROC}(f)\} = \text{intercept} + \text{slope} \times \text{logit}(f)$.

`interval(`a b n_p`)` specifies the FPR interval (a,b) and number of points (n_p) in the interval over which the ROC-GLM is to be fit. The default is `interval(0 1 10)`.

Standardization method

`pvcmeth(`*method*`)` specifies how the PVs are to be calculated. *method* can be one of the following:

`empirical`, the default, uses the empirical distribution of the test measure among controls ($D = 0$) as the reference distribution for the calculation of case PVs. The PV for the case measure y_i is the proportion of control measures $Y_{\overline{D}} < y_i$.

`normal` models the test measure among controls with a normal distribution. The PV for the case measure y_i is the standard normal c.d.f. of $(y_i - \text{mean})/\text{sd}$, where the mean and the standard deviation are calculated by using the control sample.

`tiecorr` indicates that a correction for ties between case and control values is included in the empirical PV calculation. The correction is important only in calculating summary indices, such as the AUC. The tie-corrected PV for a case with the marker value y_i is the proportion of control values $Y_{\overline{D}} < y_i$ plus one half the proportion of control values $Y_{\overline{D}} = y_i$, where $Y_{\overline{D}}$ denotes controls. By default, the PV calculation includes only the first term, i.e., the proportion of control values $Y_{\overline{D}} < y_i$. This option applies only to the empirical PV calculation method.

Covariate adjustment

`adjcov(`*varlist*`)` specifies the variables to be included in the adjustment.

`adjmodel(`*model*`)` specifies how the covariate adjustment is to be done. *model* can be one of the following:

`stratified` PVs are calculated separately for each stratum defined by *varlist* in `adjcov()`. This is the default if `adjmodel()` is not specified and `adjcov()` is. Each case-containing stratum must include at least two controls. Strata that do not include cases are excluded from calculations.

`linear` fits a linear regression of the marker distribution on the adjustment covariates among controls. Standardized residuals based on this fitted linear model are used in place of the marker values for cases and controls.

Sampling variability

`nsamp(#)` specifies the number of bootstrap samples to be drawn for estimating sampling variability of parameter estimates. The default is `nsamp(1000)`.

`nobstrap` omits bootstrap sampling and estimation of standard errors and CIs. If `nsamp()` is specified, `nobstrap` will override it.

`noccsamp` specifies that bootstrap samples be drawn from the combined sample rather than sampling separately from cases and controls; case–control sampling is the default.

`nostsamp` draws bootstrap samples without respect to covariate strata. By default, samples are drawn from within covariate strata when stratified covariate adjustment is requested via the `adjcov()` and `adjmodel()` options.

`cluster(`*varlist*`)` specifies variables identifying bootstrap resampling clusters. See the `cluster()` option of the `bootstrap` command ([R] **bootstrap**).

`resfile(`*filename*`)` creates a Stata file (a `.dta` file) with the bootstrap results for the ROC-GLM. The Stata file is called *filename*`.dta` if a single marker is specified or *filename*`#.dta` for the #th marker if more than 1 marker is included in *test_varlist*. `bstat` can be run on this file to view bootstrap results again.

`replace` specifies that if the specified file already exists, then the existing file should be overwritten.

`level(#)` specifies the confidence level for CIs as a percentage. The default is `level(95)` or as set by `set level`.

3.4 Saved results

Parameter estimates from the ROC-GLM curve fit and the corresponding bootstrap covariance matrix are available as `bootstrap postestimation` results. See also `help postest` and `help estat bootstrap`. If more than one variable is included in *test_varlist*, estimation results for the #th marker are stored under the name *rocreg_m#*. Returned estimation result matrices include the following:

Matrices

`e(b)`	1 x k matrix of ROC-GLM parameter estimates; $k = 2$ + number of covariates included in the intercept and slope terms. Columns correspond to α_0 and α_1 parameters plus coefficients for any specified covariates.
`e(V)`	k x k bootstrap covariance matrix for the k ROC-GLM parameters.
`e(GLMparm)`	n x k matrix of ROC-GLM parameter estimates. Rows correspond to the marker variables included in *test_varlist*, and columns are as for `e(b)`. Returned whether bootstrap sampling is specified or not (`nobstrap`).

(*Continued on next page*)

3.5 Example

The `rocreg` command applied to the audiology data produces the following results:

```
. use http://labs.fhcrc.org/pepe/book/data/nnhs2, clear
(Norton - neonatal audiology data)
. set seed 56930
. rocreg d y1, adjcov(currage gender) adjmodel(linear) regcov(currage)
> cluster(id) noccsamp
            ROC regression for markers: DPOAE 65 at 2kHz
       model intercept term covariates: currage
         percentile value calculation
                                method: empirical
                          tie correction: no
  Covariate adjustment for p.v. calculation:
                                method: linear model
                             covariates: currage
                                         Gender
           GLM fit of binormal curve
                     number of points: 10
                      on FPR interval: (0,1)
                        link function: probit
 model coefficient bootstrap se´s and CI´s based on sampling
  w/o respect to case/control status
          number of bootstrap samples: 1000
*****************************
 model results for marker: DPOAE 65 at 2kHz
  covariate adjustment - linear model, controls only
```

Source	SS	df	MS		
Model	2418.56541	2	1209.2827	Number of obs	= 4907
Residual	294662.363	4904	60.0861263	F(2, 4904)	= 20.13
				Prob > F	= 0.0000
Total	297080.929	4906	60.5546125	R-squared	= 0.0081
				Adj R-squared	= 0.0077
				Root MSE	= 7.7515

y1	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
currage	-.2032456	.0323905	-6.27	0.000	-.2667455	-.1397458
gender	.2471744	.2229119	1.11	0.268	-.1898327	.6841815
_cons	-1.486659	1.288611	-1.15	0.249	-4.012913	1.039596

```
************
```

```
   ROC-GLM model
Bootstrap results                               Number of obs      =      5056
                                                Number of clusters =      2741
                                                Replications       =      1000
```

	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
alpha_0	-1.2725052	-.0571745	1.0770327	-3.38345	.8384401	(N)
				-3.509356	.7178385	(P)
				-3.487457	.7813575	(BC)
alpha_1	.93723935	.0127611	.07467309	.7908828	1.083596	(N)
				.8079086	1.101941	(P)
				.7928988	1.083512	(BC)
currage	.04482277	.0016014	.02804926	-.0101528	.0997983	(N)
				-.007932	.1033131	(P)
				-.0102905	.101021	(BC)

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
```

4 Evaluating incremental value

4.1 Motivation and concept

Another way of incorporating covariate information is by evaluating incremental value. When Z is a set of known risk factors or other baseline predictors, an obvious question concerns the improvement in classification accuracy associated with adding Y to Z. Within this framework, Z is allowed to help in discriminating between cases and controls. This is in contrast to covariate adjustment methods, which characterize the classification accuracy of Y conditional on Z.

Incremental value is quantified by comparing the ROC curve for (Y, Z) to the ROC curve for Z alone. The optimal combination of Y and Z for classification is the risk score, $P(D = 1 \,|\, Y, Z)$ (McIntosh and Pepe 2002). The risk score can be estimated with a wide variety of binary regression techniques, including logistic regression, logic regression, classification trees, neural networks, and support vector machines.

(*Continued on next page*)

4.2 Estimation

We propose the following approach to estimating incremental value. First, we fit logistic regression models with and without the marker, Y:

$$P(D = 1 \,|\, Y, Z) = g(\beta_0 + \beta_1 Y + \beta_2 Z + \beta_3 Z \times Y)$$

and

$$P(D = 1 \,|\, Z) = g(\gamma_0 + \gamma_1 Z)$$

where $g(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ is the logistic function. Forms other than linear can be employed for the predictors (e.g., splines), and interactions may or may not be included. The primary advantage of using logistic regression is that the linear predictors, $g^{-1}\{P(D = 1 \,|\, Y, Z)\}$ and $g^{-1}\{P(D = 1 \,|\, Z)\}$, which have the same ROC curves as the risk scores, are consistently estimated (up to constants) with case–control data (Prentice and Pyke 1979).

Having fit the two regression models, we next calculate the associated predicted values for all the subjects in the dataset. We analyze the predicted values on the linear predictor scale, where distributional assumptions are more easily verified, noting again that the ROC curves for $g^{-1}\{P(D = 1)\}$ and $P(D = 1)$ are the same.

The final step is to plot and compare the ROC curves for the linear predictions from the two models. This can be accomplished by using `roccurve` and `comproc` (Pepe, Longton, and Janes 2009).

This procedure is simplistic in at least two respects. First, fitting and evaluating models on the same data is known to produce overly optimistic estimates of model performance. Cross-validation could be used to correct for this overoptimism. Second, the bootstrapping implemented in `roccurve` and `comproc` conditions on the fitted regression models. This bootstrapping accounts for uncertainty in the ROC estimates but not in the predicted values. Bootstrapping the entire process, from sampling to risk-score estimation to ROC estimation, would be desirable. For simplicity, we ignore these issues here but plan to implement a Stata program in the future that incorporates cross-validation and bootstrapping of the model-fitting process.

4.3 Example

We again use the audiology data to illustrate the estimation of incremental value. We evaluate the incremental value of the marker DPOAE over and above age and gender. Figure 6 shows ROC curves for two fitted logistic regression models, one including age and gender, and the other including age, gender, and DPOAE. All the covariates are modeled linearly. The ROC curves are estimated empirically (without adjustment for any covariates). We see that the inclusion of DPOAE substantially improves the ability of the model to discriminate between hearing-impaired and -unimpaired ears. The commands used to generate the estimators shown are

```
. logit d currage gender
. predict p1
. logit d currage gender y1
. predict p2
. roccurve d p1 p2, roc(0.2) cluster(id) noccsamp
```

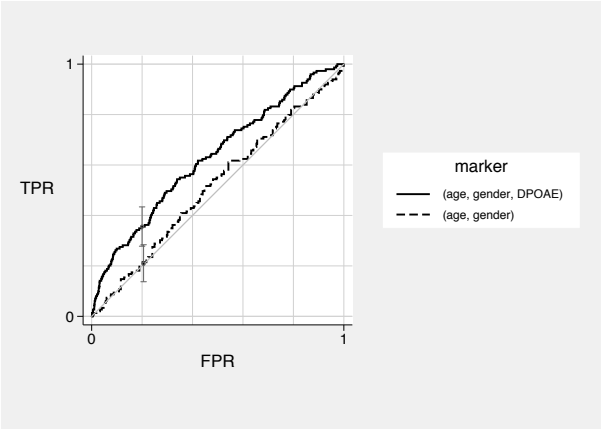

Figure 6. The incremental value of the marker DPOAE over and above age and gender, estimated with the Norton, Wang, and Ai (2004) audiology data. ROC curves are estimated for disease risk-prediction models with and without DPOAE; both models include age and gender.

5 Remarks

The methods and Stata programs presented here facilitate incorporating covariates into ROC analysis in three distinct ways: by characterizing the performance of the marker conditional on covariates (i.e., by using covariate adjustment), by allowing the accuracy of the marker to depend on the covariates (i.e., by using ROC regression), and by examining the improvement in classification accuracy associated with adding the marker to the covariates (i.e., by using the incremental value approach). The representation of the ROC curve as the c.d.f. of standardized case marker observations provides a natural means of incorporating covariate information and gives rise to parametric, semiparametric, and nonparametric estimates of the quantities of interest.

We have focused on continuous markers, but these methods can also be applied to ordinal markers.

(*Continued on next page*)

6 References

Alonzo, T. A., and M. S. Pepe. 2002. Distribution-free ROC analysis using binary regression techniques. *Biostatistics* 3: 421–432.

Baker, S. G. 2003. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute* 95: 511–515.

Dodd, L. E., and M. S. Pepe. 2003. Partial AUC estimation and regression. *Biometrics* 59: 614–623.

Dorfman, D. D., and E. Alf Jr. 1969. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. *Journal of Mathematical Psychology* 6: 487–496.

Heagerty, P. J., and M. S. Pepe. 1999. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in U.S. children. *Applied Statistics* 48: 533–551.

Janes, H., and M. S. Pepe. 2008. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. *American Journal of Epidemiology* 168: 89–97.

———. Forthcoming. Adjusting for covariate effects on classification accuracy using the covariate-adjusted ROC curve. *Biometrika*.

McIntosh, M. W., and M. S. Pepe. 2002. Combining several screening tests: Optimality of the risk score. *Biometrics* 58: 657–664.

Norton, E. C., H. Wang, and C. Ai. 2004. Computing interaction effects and standard errors in logit and probit models. *Stata Journal* 4: 154–167.

Ogilvie, J. C., and C. D. Creelman. 1968. Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology* 5: 377–391.

Pepe, M. S. 2000. An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* 56: 352–359.

Pepe, M. S., and T. Cai. 2004. The analysis of placement values for evaluating discriminatory measures. *Biometrics* 60: 528–535.

Pepe, M. S., R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui. 2001. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 93: 1054–1061.

Pepe, M. S., and G. Longton. 2005. Standardizing diagnostic markers to evaluate and compare their performances. *Epidemiology* 16: 598–603.

Pepe, M. S., G. Longton, and H. Janes. 2009. Estimation and comparison of receiver operating characteristic curves. *Stata Journal* 9: 1–16.

Prentice, R., and R. Pyke. 1979. Logistic disease incidence models and case–control studies. *Biometrika* 66: 403–411.

About the authors

Margaret Pepe is a full member, Gary Longton is a statistical research associate, and Holly Janes is an assistant member in the Public Health Sciences Division of the Fred Hutchinson Cancer Research Center in Seattle. A focus of their research is on the development of new methodology for diagnostic tests and biomarkers, with support provided by the National Cancer Institute (CA 129934 and CA 086368) and the National Institute for General Medical Studies (GM 054438). Pepe, Longton, and Janes also teach courses on statistical methods for evaluating tests and biomarkers.