3-1-2006

Evaluating the Predictiveness of a Continuous Marker

Ying Huang
University of Washington, ying@u.washington.edu

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Ziding Feng
University of Washington & Fred Hutchinson Cancer Research Center, zfeng@fhcrc.org

Evaluating the Predictiveness of a Continuous Marker

Ying Huang, Margaret Sullivan Pepe† and Ziding Feng

Fred Hutchinson Cancer Research Center Public Health Sciences

1100 Fairview Avenue N., M2-B500, Seattle, WA 98109-1024

and

University of Washington Biostatistics Department F-600 Health Sciences Building

Box 357232, Seattle, WA 98195-7232

†*Corresponding author's address:* mspepe@u.washington.edu

Summary.

Consider a continuous marker for predicting a binary outcome. For example, serum concentration of prostate specific antigen (PSA) may be used to calculate the risk of finding prostate cancer in a biopsy. In this paper we argue that the predictive capacity of a marker has to do with the population distribution of risk given the marker and suggest a graphical tool, the predictiveness curve, that displays this distribution. The display provides a common meaningful scale for comparing markers that may not be comparable on their original scales. Some existing measures of predictiveness are shown to be summary indices derived from the predictiveness curve. We develop methods for making inference about the predictiveness

curve, for making pointwise comparisons between two curves and for evaluating covariate effects. Applications to risk prediction markers in cancer and cystic fibrosis are discussed.

KEY WORDS: risk, classification, explained variation, biomarker, ROC curve, prediction

1. Background

The Early Detection Research Network (EDRN) is a national network of biomarker development laboratories and clinical centers sponsored by the National Cancer Institute (Vastag 2000). Two of its goals are to develop biomarkers for cancer screening and for cancer risk prediction. The statistical focus in cancer screening is on the capacity of a marker to accurately classify subjects as diseased or not (Pepe et al 2001). Classification performance parameters such as sensitivity and specificity are of key interest, because ultimately it is the proportion of diseased subjects detected (sensitivity) and the proportion of non-diseased subjects unnecessarily referred for work-up (1-specificity) that enter into decisions about screening policy. The evaluation of markers for cancer risk prediction, however, requires a different approach. In this context we need to quantify how well a marker identifies people at high or low risk for cancer. A cancer risk prediction marker might be used to select subjects for a prevention intervention or indeed for screening, but does not classify subjects directly. Indeed Gail and Pfeiffer (2005) note that criteria for cancer risk prediction markers will often be much less stringent than those required of screening markers.

In this paper we propose a graphical display to aid in the assessment of risk prediction markers. The *predictiveness curve* shows the predictive capacity of a marker. An important attribute is that it provides a common scale for comparing risk prediction markers. The predictiveness curve has been suggested previously by Bura and Gastwirth (2001) and Copas (1999), albeit with different terminology. However, their focus was on inference for summary indices. They did not address inference for the curve itself or comparisons between curves, nor

did they consider curves for subpopulations defined by covariates. In this paper, in addition to addressing these points, we demonstrate the practical usefulness of the predictiveness curve in assessing the value of risk prediction markers. We illustrate the methodology using two datasets. The first is a non-cancer application. It concerns major pulmonary infections in children with cystic fibrosis and the capacity of measures of lung function and nutritional status to predict them. The second concerns prostate specific antigen (PSA), a widely used biomarker for prostate cancer.

2. Predictiveness of a Binary Marker

Let D denote the binary outcome and denote the marker by Y. The risk associated with marker value $Y = y$ is

$$risk(y) \equiv P(D = 1|Y = y).$$

Throughout most of this paper we assume that larger values of Y are associated with increasing risk, but generalize the ideas in Section 9. Although our interest is primarily in evaluating continuous markers, we first consider the simple setting when the marker is binary. In that case, subjects either have the lower risk level, $risk(0) = P(D = 1|Y = 0)$, or the higher value, $risk(1) = P(D = 1|Y = 1)$.

Frequently the relative risk is used to summarize the predictiveness of a marker. However, clearly the absolute levels of risk, not just their ratio, are important in describing the predictive capacity of a marker. For example, a marker with relative risk equal to 10 may correspond to absolute risks of $(risk(0) = 0.1\%, risk(1) = 1\%)$ or $(risk(0) = 1\%, risk(1) = 10\%)$. These two scenarios have very different implications if risks below 5% say are considered unimportant and risks above 5% are considered "high."

Reporting of absolute risks, however, is not sufficient either. The proportions of the population who attain the lower and higher risk levels are also crucial components of predictiveness. For example, when seeking a marker that identifies subjects at "high risk", a

marker with larger high risk prevalence, $P[Y = 1]$, may be preferable even if the absolute high risk value, $risk(1)$, is somewhat smaller.

The definitions of "high" and "low" risk depend entirely on the clinical context and include consideration of overall prevalence, consequences of disease, and interventions for subjects in the low and high risk strata. In general, once appropriate thresholds have been decided upon, one marker is preferable to another if it leaves fewer subjects in the intermediate equivocal risk range.

In conclusion, to fully describe the predictiveness of a binary marker, we stipulate that one should report the absolute risks and the frequencies of those risks in the population. That is, one should report the distribution of risk in the population conferred by the marker. The remainder of this paper expands on this theme in relation to continuous markers.

3. The Predictiveness Curve

We propose the predictiveness curve, $R(v)$ versus v, for describing the predictive capacity of a marker where $R(v)$ is the risk associated with the v^{th} quantile of the marker:

$$R(v) = P[D = 1|Y = F^{-1}(v)],$$

and F is the cumulative distribution function of the marker. Figure 1 displays curves for $-\text{FEV}_1$, a measure of lung function and $-weight$, a measure of nutritional status, for predicting serious lung infection in the following year among cystic fibrosis patients. The markers are negated to conform to our convention that increasing values are associated with increasing risk. Details of the data will be discussed in Section 7.

Observe that by using the scale $v = F(Y)$ on the x-axis, the markers are transformed to a common scale. This facilitates their comparison, whereas on their original scales the markers are not comparable. We see for example that at the 90^{th} percentile of $-\text{FEV}_1$, the risk is 0.76, whereas the risk is only 0.58 at the 90^{th} percentile of $-weight$. Patients in the top 10% of the marker distribution are at greater risk when lung function rather than nutritional

status is used as the risk prediction marker. Pulmonary function is also a better marker of low risk. The bottom 10% have risks in the range (0.25, 0.28) according to $-weight$ but in a much lower range, (0.01,0.15) according to $-\text{FEV}_1$.

Another way of looking at the predictiveness curve is to consider the inverse function. We see that $R^{-1}(p)$ is the proportion of the population with risks less than or equal to p. Suppose p_L is a threshold that defines 'low risk' and p_H is a threshold that defines 'high risk.' Then the proportions of the population with low, high, and equivocal risks are $R^{-1}(p_L), 1-R^{-1}(p_H)$ and $R^{-1}(p_H)-R^{-1}(p_L)$, respectively. To illustrate in the cystic fibrosis example, suppose we take $p_H = 0.75$ and $p_L = 0.25$, then lung function is predictive of low risk in $R^{-1}(0.25) = 29\%$ of the population, of high risk in $1 - R^{-1}(0.75) = 12\%$ and it leaves 59% of patients in the equivocal risk range. Nutritional status on the other hand is completely uninformative about high or low risk. Knowledge of weight leaves all 100% of patients with risks in the equivocal risk range, (0.25,0.75).

Interestingly, the predictiveness curve displays the distribution of $risk(Y)$ in the population. As mentioned above, and most easily seen from the plot itself, $R^{-1}(p)$ is the proportion of the population with risks less than p. Mathematically we write

$$R^{-1}(p) = P[risk(Y) < p]$$

is the cumulative distribution of $risk(Y)$. Correspondingly $R(v)$ is the $100 \times v^{th}$ percentile of $risk(Y)$ in the population. We find the display simple and useful for describing the predictive capacity of a marker. It conveys the essential elements of our concept of predictiveness and leads us to a general definition for the predictiveness of a marker in the population.

Definition. Predictiveness of $Y \equiv$ the distribution of $risk(Y)$

A marker that is uninformative about risk assigns equal risk to all subjects, $risk(Y) = P[D = 1|Y] = P[D = 1] = \rho$. The corresponding predictiveness curve is the horizontal line

at $R(v) = \rho$, $v \in (0,1)$. On the other hand, a perfect marker assigns $risk(Y) = 1$ for the proportion ρ of subjects with $D = 1$ and $risk(Y) = 0$ for the proportion $1 - \rho$ with $D = 0$. Correspondingly, its predictiveness curve is the step function $R(v) = I[(1-\rho) < v]$, where $I[.]$ is the indicator function. Most risk prediction markers are imperfect, lying between these extremes. Better markers have steeper curves corresponding to wider variation in risk. Note that $\int R(v)dv = \int P[D = 1|Y]dF(Y) = \rho$. Therefore $\int (R(v) - \rho)dv = 0$, which implies that $\int_0^{v^*}(\rho - R(v))dv = \int_{v^*}^1 (R(v)) - \rho)dv$ where v^* is the point where $R(v^*) = \rho$. In other words, the areas between the curve and the horizontal line at ρ that are above and below the horizontal line are equal. The horizontal line at ρ serves as a useful benchmark and visual aid in evaluating predictiveness curves.

4. Estimation

We now turn to the task of estimating the predictiveness curve. Suppose data from a random sample of n independent identically distributed subjects are available, $\{(Y_i, D_i), i = 1, \ldots, n\}$. We model the risk as a parametric increasing function of Y into (0,1):

$$P[D = 1|Y] = G(\beta, Y)$$

where G has the form of a cumulative distribution function (cdf). Assume that an asymptotically normal estimator of β is employed with $\text{var}(\sqrt{n}(\hat{\beta} - \beta)) = \Sigma(\beta)$. For example, $\hat{\beta}$ might be the maximum likelihood estimate from a linear logistic model $G(\beta, Y) = \exp(\beta_0 + \beta Y)/\{1 + \exp(\beta_0 + \beta Y)\}$. We prefer to employ more flexible models and use the 3 parameter Box-Cox family (Cole and Green, 1992) in our illustrations. Let $\hat{F}$ be the empirical cdf of Y. Then

$$\hat{R}(v) \equiv G(\hat{\beta}, \hat{F}^{-1}(v)) \qquad v\epsilon(0,1).$$

Theorem 1

The asymptotic distribution of $\sqrt{n}(\hat{R}(v) - R(v))$ is mean 0, normal, with variance $\sigma^2(v)$ where

$$\sigma^2(v) = \left(\frac{\partial R(v)}{\partial \beta}\right)^T \Sigma(\beta) \left(\frac{\partial R(v))}{\partial \beta}\right) + v(1-v)\left(\frac{\partial R(v)}{\partial v}\right)^2 \quad (1)$$

and $0 < v < 1$.

■

The result indicates that the variance of $\hat{R}(v)$ is comprised of two additive components. The first is due to variability in $\hat{\beta}$ while the second is due to variability in $\hat{F}^{-1}(v)$. Observe that the magnitude of the second component depends on the slope of the predictiveness curve at v. The variability due to $\hat{F}^{-1}(v)$ is more important when $R(v)$ is steep. It makes sense intuitively that imprecision in $\hat{F}^{-1}(v)$ will have a greater impact on $\hat{R}(v)$ when $(\partial/\partial v)R(v)$ is larger. Asymptotic theory for the inverse function is provided in the next result.

Theorem 2

$\sqrt{n}(\hat{R}^{-1}(p) - R^{-1}(p))$ has an asymptotically normal distribution with mean 0 and variance $\tau^2(p)$ where

$$\tau^2(p) = \left[\frac{\partial R^{-1}(p)}{\partial \beta}\right]^T \Sigma(\beta) \left[\frac{\partial R^{-1}(p)}{\partial \beta}\right] + R^{-1}(p)(1 - R^{-1}(p)) \quad (2)$$

and p is in the range of $\{R(v) : v\epsilon(0,1)\}$

■

The variance again is comprised of two additive components, one due to $\hat{\beta}$ and the other due to $\hat{F}$. If β were known precisely, then $\hat{R}^{-1}(p)$ is the binomial proportion of subjects with $risk(Y) < p$, so its variance is the binomial variance $R^{-1}(p)(1 - R^{-1}(p))/n$ as indicated by (2). On the other hand, if F, the population distribution of Y, were known, then $\hat{R}^{-1}(p) = F(G^{-1}(\hat{\beta}, p))$, which only includes $\hat{\beta}$ as a random variable, and its variance is given by the first component of (2). Observe also the simple relationship between $\tau^2(p)$

and $\sigma^2(v)$ at $v = R^{-1}(p)$: $\sigma^2(v) = ((\partial/\partial v)R(v))^2\tau^2(p)$. Variability in $\hat{R}^{-1}(p)$ is magnified by $(\partial/\partial v)R(v)$ on the scale for $\hat{R}(v)$.

Consistent estimates of $\sigma^2(v)$ and $\tau^2(p)$ are obtained by substituting estimated quantities into expressions (1) and (2). Derivatives with respect to β of $\hat{R}(v) = G(\beta, \hat{F}^{-1}(v))$ and $\hat{R}^{-1}(p) = \hat{F}(G^{-1}(\hat{\beta}, p))$ are easily obtained since G and G^{-1} are simple differentiable functions of β. To estimate $(\partial/\partial v)(R(v)) = [(\partial/\partial y)G(\beta, y)][(\partial/\partial v)F^{-1}(v)]$ where $y = F^{-1}(v)$, the first component is straightforward while the second involves the density of Y, $(\partial/\partial v)F^{-1}(v) = 1/f(F^{-1}(v))$. We use a Gaussian kernel density estimate for f in our applications with bandwidth h optimal for normally distributed data (Silverman, 1986):$h = n^{-\frac{1}{5}}1.06\,\min(\text{standard deviation, interquartile range}/1.349)$.

5. Simulation Studies

We performed a limited simulation study to investigate the use of large sample inference for $R(v)$ and $R^{-1}(p)$ in finite sample studies. Data were simulated according to two models. In the first, F is a standard normal distribution and the risk function is linear logistic, $G(\beta, Y) = \exp(\beta_0 + \beta_1 Y)/(1 + \exp(\beta_0 + \beta_1 Y))$. We chose (β_0, β_1) such that $R(0.1)$ and $R(0.9)$ were at pre-defined values. In the second, F is standard log normal and the risk function is from the Box-Cox family: $G(\beta, Y) = \Phi(\beta_0 + \beta_1 Y^{(\beta_2)})$ where $Y^{(\lambda)} = (Y^\lambda - 1)/\lambda$ when $\lambda \neq 0$ and $Y^{(\lambda)} = \log Y$ when $\lambda = 0$. In this setting we fixed $\beta_2 = -0.4$, 0, or 0.4 and then chose (β_0, β_1) based on specified values for $R(0.1)$ and $R(0.9)$. We fit the predictiveness curves using the correct form for $G(\beta, Y)$. That is, we did not investigate the robustness of these models to misspecification. In practice one should check for goodness of fit, investigating multiple model forms if necessary.

Tables 1 and 2 show the results of our simulations for one of the Box-Cox model simulations. We found that bias was minimal. Variance estimates reflected the actual sampling variability with sample sizes of $n = 500$ or more. Consequently the coverage of 95% confi-

dence intervals was excellent in moderate to large samples but lower than the nominal level with $n = 100$. Problems occurred only at the edges of the predictiveness curve, at $v = 0.1$ and $v = 0.9$. Considering that with $n = 100$ only 10 observations lay beyond these points, the reduced performance under these circumstances seems reasonable. With larger sample sizes, ($n \geq 500$), inference based on asymptotic inference appeared to work very well across all of the scenarios we studied (additional results not shown).

6. Further Inferential Techniques

6.1 *Comparing Markers*

A key attribute of the predictiveness curve is that it provides a common relevant scale for comparing risk prediction markers. Here we consider formal comparisons between two markers. Comparisons might be based on the difference in risk percentiles, $R_A(v_0) - R_B(v_0)$, at some $v = v_0$ of interest where subscripts denote markers A and B. The Appendix summarizes asymptotic distribution theory for $\hat{R}_A(v) - \hat{R}_B(v)$, assuming that both markers are measured on the same individuals. The estimated standard error of the difference is calculated using steps analogous to those described earlier for calculating the estimated standard error of $\hat{R}(v)$. A p-value can then be based on the Z-statistic: $\hat{R}_A(v) - \hat{R}_B(v)$ divided by the estimated standard error.

A particularly compelling case can be made for comparing markers on the basis of $\hat{R}^{-1}(p)$ where p is a threshold that defines high (or low) risk. One marker of high risk would be preferred over another if it identifies a greater fraction of people at high risk. That is, if the high risk threshold is p_H and $R_A^{-1}(p_H) - R_B^{-1}(p_H) > 0$ then marker B is a better marker of high risk. Asymptotic distribution theory for $\hat{R}_A^{-1}(p) - \hat{R}_B^{-1}(p)$, provided in the Appendix can serve as the basis of confidence intervals and hypothesis testing.

Our methodology concerns pointwise comparisons between markers. The simple clinically relevant interpretations for points on the predictiveness curve and its inverse motivate making such comparisons in practice. However, statistical power might be greater for statistics based

on summary indices. Gail and Pfeiffer (2005) note that several indices of predictability are functionals of the predictiveness curve. For example, the average entropy (Shapiro 1977) is $\int\{R(v)\log R(v)+(1-R(v))\log(1-R(v))\}dv$ and the risk variance is $\int(R(v)-\rho)^2dv$. Both give rise to measures of the proportion of explained variation (Mittlebock and Schemper, 1996) which are existing summary indices of predictiveness. Bura and Gastwrith (2001) suggest the total gain, $TG = \int |R(v)-\rho|dv$, as a summary index. Methods for making formal comparisons between predictiveness curves based on summary indices, however, have not been studied yet. Bootstrapping could be used for inference in practice.

6.2 *Covariate Specific Predictiveness Curves*

The predictiveness of a marker can vary across populations. This can happen if the marker distribution varies or if there is an interaction between the marker and a covariate on risk. In addition, the usefulness of a risk prediction marker may vary with the overall risk. In a low risk population, the marker may not identify any high risk subjects while in a moderate risk population it might, even if the distribution of the marker and its association with risk are the same. In this section we consider how to estimate predictiveness curves in subpopulations defined by covariates denoted by Z.

For a discrete covariate, one can simply stratify and estimate stratum specific predictiveness curves, $R_Z(v)$, as described earlier. More generally we model the effect of Z on risk with

$$P(D=1|Y,Z) = G(\beta, Y, Z)$$

where, as before, G is monotone increasing in Y, and we use a semiparametric location-scale model (Heagerty and Pepe, 1999) for the distribution of Y given Z

$$F_z(y) = P(Y \le y|Z=z) = F_0\left(\frac{y-\mu_z}{\sigma_z}\right)$$

where $\mu_Z = \gamma' U(Z)$ and $\log(\sigma_Z) = \delta' W(Z)$, and $U(Z)$ and $W(Z)$ are specified functionals of Z. For example, for binary Z, $U(Z)$ and $W(Z)$ could be $(1, Z)$, while for continuous Z,

$U(Z)$ and $W(Z)$ could be a B-spline basis for Z. Writing $U_i = U(Z_i)$ and $W_i = W(Z_i)$, the estimators $\hat{\gamma}$ and $\hat{\delta}$ are solutions to

$$\sum_{i=1}^{n} U_i(Y_i - \gamma' U_i)/\sigma^2_{Z_i} = 0$$

$$\sum_{i=1}^{n} W_i[(Y_i - \gamma' U_i)^2 - \sigma^2_{Z_i}]/\sigma^2_{Z_i} = 0.$$

Denoting the empirical cumulative distribution of standardized residuals by $\hat{F}_0(\cdot)$, the Z=z covariate-specific marker distribution estimate is

$$\hat{F}_z(y) = \hat{F}_0\left(\frac{y - \hat{\gamma}'u}{e^{\hat{\delta}'w}}\right)$$

where $u = U(z)$ and $w = W(z)$. The corresponding v^{th} quantile is

$$\hat{F}_z^{-1}(v) = \hat{\gamma}'u + e^{\hat{\delta}'w}\hat{F}_0^{-1}(v).$$

Having fit the risk model, through maximum likelihood or otherwise, the covariate specific predictiveness curve estimate is

$$\hat{R}_z(v) = G(\hat{\beta}, \hat{F}_z^{-1}(v), z).$$

Similarly, the estimated inverse is

$$\hat{R}_z^{-1}(p) = \hat{F}_z(G^{-1}(\hat{\beta}, p, z)).$$

Asymptotic distribution theory is provided in the appendix.

7. The Cystic Fibrosis Data

Cystic fibrosis is a genetic disorder that results in impaired ion transport across the cell membrane. Its effects on pulmonary and gastrointestinal systems lead to progressive deterioration. Predicted survival currently extends into the mid 30s (Cystic Fibrosis Foundation, 2004). The main culminating event that leads to death is acute pulmonary exacerbation, i.e., lung infection requiring intravenous antibiotics.

To illustrate our methodology we use data from the Cystic Fibrosis Registry, a database maintained by the Cystic Fibrosis Foundation, containing annually updated information on over 20,000 people diagnosed with CF and living in the USA. We consider weight and FEV_1 measured in 1995 to predict pulmonary exacerbations in 1996. Data for 11,960 patients are analyzed of whom 5094 (42%) had at least one pulmonary exacerbation. Patients 6 years of age and older are included. Weight is standardized for age and gender (Hamill et al., 1977) and FEV_1 is standardized for age, gender and height (Knudson et al., 1983). See Moskowitz and Pepe (2004) for more details. Figure 1 shows the estimated predictiveness curves for $-\text{FEV}_1$ and $-weight$. We have referred to them as population curves earlier in the paper when illustrating concepts but now acknowledge their sampling variability.

Table 3 shows point estimates and confidence intervals for $R(v)$ and $R^{-1}(p)$. We have tight confidence intervals for the estimates whose values were already mentioned in relation to this curve in Section 3. Confidence intervals calculated using bootstrap resampling were almost identical to those based on asymptotic theory. The second column shows the % variance in $\hat{R}(v)$ due to $\hat{F}$, i.e., the second component of (1). Observe that it is larger for $-\text{FEV}_1$ than for $-weight$ presumably because the predictiveness curve for $-\text{FEV}_1$ is more steep. An hypothesis test based on the difference in predictiveness estimates at $v = .1$ yields p-value< 0.01. This test uses the asymptotic variance expression given in the Appendix. At $v = 0.9$ the difference in $R(v)$ for the two markers is also statistically significant (p-value < 0.01).

We next consider the predictiveness of lung function ($Y = -\text{FEV}_1$) in subpopulations defined by their nutritional status ($Z = weight$). We modeled the risk of pulmonary exacerbation with the Box-Cox form

$$P[D = 1|Y, Z] = \Phi(\beta_0 + \beta_1 Y^{(\lambda)} + \beta_2 Z)$$

and used a semiparametric location scale model for $\log(-Y)$ with mean a natural cubic spline

function of Z having knots at the 0.30 and 0.70 quantiles of Z and log standard derivation also a natural cubic spline with knot at the median of Z. Fitted predictiveness curves are shown in Figure 2 for subjects with weight at the median, first and third quartiles. Observe that the incidence of pulmonary exacerbation varies across these three populations, from 58% in subjects at the first quartile of weight to 30% in subjects at the third quartile. In the latter population, FEV_1 identifies .41 of subjects below the low risk threshold of 25% (95% CI = (0.38,0.43)) but only 0.01 with risks above the high risk threshold of 75% (95% CI=(0.01,0.02)). On the other hand, in the high risk population FEV_1 identifies only 0.06 (95% CI=(0.05,0.07)) low risk subjects and 0.23 (95% CI=(0.20,0.25)) high risk subjects. Therefore FEV_1 is a particularly useful marker of high risk in subjects at already somewhat elevated risk due to their nutritional status and similarly it is more useful as a low risk marker in subjects at somewhat reduced risk based on their weight.

8. Markers for Prostate Cancer

The Prostate Cancer Prevention Trial was a randomized prospective study of men with PSA < 3.0 ng/mL, aged 55 years and older who were followed up for 7 years with annual PSA measurements. A biopsy was recommended for all men either during or at the end of the study. Thompson et al. (2006) identified 5519 men on the placebo arm of the trial who had undergone prostate biopsy and had a PSA and digital rectal exam (DRE) during the year prior to biopsy and at least 2 PSA values from the 3 years prior to biopsy. Prostate cancer risk was evaluated as a function of PSA, PSA velocity and several other variables including age, family history, DRE and prior prostate biopsy. Here we use the data to compare PSA and PSA velocity as predictors of prostate cancer risk. 21.9% of men were found to have prostate cancer. Figure 3 top panel displays the predictiveness curves for PSA and PSA velocity. The curve for PSA velocity is more shallow, indicating that it is a poorer marker of risk for this general class of prostate cancers. The 90^{th}percentile of risk is 0.291

according to PSA velocity while it is higher, 0.369, according to the absolute most recent PSA measurement ($p < 0.01$). At the low end of the scale, the 10^{th}risk percentiles based on PSA velocity and PSA are 0.149 and 0.091 ($p < 0.01$) respectively, again suggesting that PSA is the better marker of risk. According to PSA velocity 2.9% of men can be classified as having risk below 10% while far more, 13.6%, qualify as low risk when using most recent PSA as the marker ($p < 0.01$). In addition, a greater fraction are found to have risks above 30% with PSA, $1 - \hat{R}^{-1}(.30) = 23.0\%$, than with PSA velocity $1 - \hat{R}^{-1}(0.30) = 8.0\%(p < 0.01)$.

Prostate cancer biopsy specimens are classified using the Gleason scoring system with higher scores associated with more aggressive disease. The bottom panel of Figure 3 shows the same markers as predictors of high-grade prostate cancer (Gleason score ≥ 7) which occurred in only 4.7% of men. Again, the $10^{\text{th}}$and 90^{th}percentiles of risk are better for PSA than for PSA velocity. The 10^{th}percentiles are 0.4% versus 1.3% (p=0.05) and the 90^{th}percentiles are 11.1% versus 8.5% ($p < .01$) for the two markers. The proportion of subjects with risks greater than 10% is 0.063 according to the PSA velocity marker while the proportion is much larger, 0.132, according to the absolute PSA marker ($p < 0.01$). We conclude that absolute PSA is also a better risk prediction marker than PSA velocity for high grade prostate cancer.

9. Discussion

The predictiveness curve provides a complete and conceptually simple description of the capacity of a marker to predict risk. It is not an entirely new proposal. The idea of categorizing a continuous marker according to quartiles or quintiles say and documenting the proportions of subjects with $D = 1$ in each category is not uncommon in the applied literature (see Willett et al., 1987 for example). Our proposal builds on and formalizes the idea, allowing the marker scale to remain continuous and restricting the risk function to be monotone. Our method has the added advantage that it provides inference about the inverse function,

$R^{-1}(p)$ (or $1-R^{-1}(p)$) the proportion of the population with risks below (or above) p, which are often of key interest.

Receiver operating characteristic curves generalize to continuous markers the notions of sensitivity, $P(Y = 1|D = 1)$, and specificity, $P(Y = 0|D = 0)$, that are defined for binary markers. The idea is to use a threshold c to define a series of binary markers, "$Y > c$", and to plot the corresponding sensitivity versus 1−specificity for all values of c. In a similar vein, Moskowitz and Pepe (2004) proposed generalizing the binary marker notions of positive and negative predictive values, PPV $= P(D = 1|Y = 1)$ and NPV $= P(D = 0|Y = 0)$, to continuous markers using thresholds. They proposed the positive predictive value (PPV) curve, which is a plot of PPV$(v) = P(D = 1|Y > F^{-1}(v))$ versus v. These curves are mathematically related to the predictiveness curve

$$\text{PPV}(v) = \int_v^1 R(u)du/(1-v).$$

In words, the PPV of the decision rule $F(Y) > v$ is the average $risk(Y)$ for Y that satisfy the positivity criterion. We note that the PPV curve is concerned with the performance of classification rules, not directly with predictiveness. An estimated predictiveness curve gives rise to an estimate of the predictive value curve. However, in contrast the empirical nonparametric methods used by Moskowitz and Pepe (2004) to estimate the PPV curve do not give rise to a simple estimate of $R(v)$. Numerical derivatives are required and mononicity cannot be guaranteed. Observe that neither ROC curves nor positive predictive value curves show the population distribution of risk that is displayed by predictiveness curves.

A key element of our proposal is that $risk(Y)$ is monotone increasing. This allows us to interpret $R^{-1}(p)$ as the proportion of the population with risks below p, and to say that a proportion v of the population have risks below $R(v)$. Some markers may not have monotone increasing risk functions. To accommodate such markers a more general definition of the

predictiveness curve can be provided:

$$R(v) = p : P[risk(Y) \leq p] = v \tag{3}$$

That is, $R(v)$ is the $100 \times v^{th}$ percentile of $risk(Y)$ in the population. Equivalently, $R^{-1}(p)$ is the cdf of the random variable $risk(Y)$. The ordering on the x-axis is according to $risk(Y)$ and is equivalent to that based on Y if $risk(Y)$ is a monotone increasing function. For the more general case, one can fit a risk model that allows non-monotonicity and plot the empirical percentiles for $\hat{R}(v)$.

In most applications however, we expect that the risk function is monotone increasing with the marker. Efficiency is likely gained by incorporating this restriction into the model. We used the Box-Cox family of distributions to fit monotone predictiveness curves to data. This 3 parameter family is reasonably flexible. It includes a wide variety of shapes (Cole and Cole 1992). However, further research is required to determine if another parametric family would be preferable. For fitting non-monotone curves we prefer B-splines because of their local nature and numerical stability. However analogues for fitting monotone curves do not appear to be available.

We have considered simple continuous prediction markers in this paper. However Y could be a function of multiple markers and risk factors. It might be a clinical prediction score derived from fitting a risk model to data. We note that if the risk score is to be evaluated with predictiveness curves estimated from the same data as used to develop the score, then issues pertaining to shrinkage, (ie., overoptimism of $\hat{R}(v)$ estimated from the same data used to develop the linear combination) would need to be addressed. We leave that for future research.

Acknowledgements

This work was supported in part by NIH grants UO1 CA086368 and RO1 GM54438. We are grateful to colleagues at the University of Washington and Fred Hutchinson Cancer Research Center for their comments on this methodology and to Gary Longton and Noelle Noble for assistance with manuscript preparation. We thank investigators of the PCPT and of the Cystic Fibrosis Foundation for providing data used to illustrate the methods.

Résumé

References

Bura E., and Gastwirth, J. L. (2001) The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually. *Biometrical Journal* **43**(1), 5–21.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine* **11**, 1305–19.

Copas, J. B. (1999). The effectiveness of risk scores: the logit rank plot. *Journal of the Royal Statistical Society C, Applied Statistics* **48**, 165–183.

Cystic Fibrosis Foundation. Patient Registry 2004 Annual Report, Bethesda, MD, USA.

Gail, M. H. and Pfeiffer, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6**(2), 227–39.

Hamill, P. V., Drizd, T. A., Johnson, C. L., Reed, R. B. and Roche, A. F. (1977). NCHS growth curves for children birth-8 years. United States, Vital Health Statistics 11, pp. 1-74. Washington, DC.

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

Knudson, R. J., Lebowitz, M. D., Holberg, C. J. and Burrows, B. (1983). Changes in the normal maximal expiratory flow-volume curve with growth and aging. *American Review*

of Respiratory Disease **127**, 725-734.

Mittlebock, M. and Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine* **15**:1987–1997.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113–127.

Pepe, M. S ., Etzioni, R., Feng, Z., et al. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**(14), 1054–1061.

Shapiro, A.R. (1977). The evaluation of medical predictions. *New England Journal of Medicine* **296**, 1509–1514.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Thompson, I. M., Pauler Ankerst, D., Chi, C., et al. (2006). Screen-based prostate cancer risk: Results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute* **98**, 529–534.

Vastag, B. (2000). Detection network gives early cancer tests a push. *Journal of the National Cancer Institute* **92**:786–788

Willett, W.C., Stampfer, M.J., Colditz, G.A., Rosner, B.A., Hennekens, C.H. and Speizer, F.E. (1987). Dietary fat and the risk of breast cancer. *New England Journal of Medicine* **316**(1):22–28

Appendix: Large Sample Distributions

Let

$$\begin{aligned}\Psi_{1i} &= -\frac{\partial R(v)}{\partial v}\left(I\left[Y_i \leq F^{-1}(v)\right] - v\right)\\ \Psi_{2i} &= I\left[G(\beta, Y_i) \leq p\right] - P\left[G(\beta, Y) \leq p\right] = I\left[Y_i \leq G^{-1}(\beta, p)\right] - R^{-1}(p)\end{aligned}$$

and assume $\sqrt{n}(\hat{\beta}_n - \beta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \Psi_{3i} + o_p(1)$, where $\Psi_{3i}, i = 1, ..., n$ are independently

identically distributed variables with $E(\Psi_{3i}|Y_i) = 0$ and $\text{var}(\Psi_{3i}) \equiv \Sigma(\beta)$. Observe that Ψ_{1i} and Ψ_{2i} are also mean zero random variables with variances $v(1-v)\left(\frac{\partial R(v)}{\partial v}\right)^2$ and $R^{-1}(p)(1-R^{-1}(p))$, respectively. When $v = R^{-1}(p)$, $\Psi_{1i} = -\frac{\partial R(v)}{\partial v}\Psi_{2i}$.

In the supplementary appendix we show

$$\sqrt{n}\left(\hat{R}(v) - R(v)\right) = n^{-1/2}\left(\frac{\partial R(v)}{\partial \beta}\right)\sum_{i=1}^{n}\Psi_{3i} + n^{-1/2}\sum_{i=1}^{n}\Psi_{1i} + o_p(1)$$

$$\sqrt{n}\left(\hat{R}^{-1}(p) - R^{-1}(p)\right) = n^{-1/2}\left(\frac{\partial R^{-1}(p)}{\partial \beta}\right)\sum_{i=1}^{n}\Psi_{3i} + n^{-1/2}\sum_{i=1}^{n}\Psi_{2i} + o_p(1)$$

A.1 Comparing Markers

Subscripts A and B are used to denote the markers.

Result A.1

The asymptotic distribution of $\sqrt{n}(\hat{R}_A(v) - \hat{R}_B(v) - (R_A(v) - R_B(v)))$ is normal with mean 0 and variance

$$\text{var}\left(\frac{\partial R_A(v)}{\partial \beta_A}\Psi_{3A} - \frac{\partial R_B(v)}{\partial \beta_B}\Psi_{3B}\right) + \text{var}\left(\Psi_{1A} - \Psi_{1B}\right)$$

where β_A and β_B are the parameters in the risk models for $P(D = 1|Y_A)$ and $P(D = 1|Y_B)$ respectively.

■

Result A.2

The asymptotic distribution of $\sqrt{n}(\hat{R}_A^{-1}(p) - \hat{R}_B^{-1}(p)) - (R_A^{-1}(p) - R_B^{-1}(p))$ is normal with mean 0 and variance

$$\text{var}\left(\frac{\partial R_A^{-1}(p)}{\partial \beta_A}\Psi_{3A} - \frac{\partial R_B^{-1}(p)}{\partial \beta_B}\Psi_{3B}\right) + \text{var}(\Psi_{2A} - \Psi_{2B})$$

Asymptotic variances can be estimated by substituting empirical or estimated quantities as necessary.

A.2 Covariate Specific Predictiveness Curves

Result A.3

Since $\hat{\gamma}$ and $\hat{\delta}$ are solutions to estimating equations we can write them as $\sqrt{n}\left(\hat{\gamma}-\gamma\right) \equiv n^{-1/2}\sum_{i=1}\gamma_i$ and $\sqrt{n}\left(\hat{\delta}-\delta\right) \equiv n^{-1/2}\sum_{i=1}\delta_i$.

Let

$$\begin{aligned}\Psi_{1iz} &= -\frac{\partial R_z(v)}{\partial v}\left(I\left(\frac{Y_i-\gamma' U_i}{e^{\delta' W_i}} \le F_0^{-1}(v)\right)-v\right) \\ &+ \frac{\partial G(\beta, F_z^{-1}(v), z)}{\partial F_z^{-1}(v)}\left(u - e^{\delta' w}E(U/e^{\delta' W})\right)'\gamma_i \\ &+ \frac{\partial G(\beta, F_z^{-1}(v), z)}{\partial F_z^{-1}(v)}(F_z^{-1}(v)-\gamma' u)\left(w-E(W)\right)'\delta_i\end{aligned}$$

$\sqrt{n}(\hat{R}_z(v)-R_z(v))$ and $\sqrt{n}(\hat{R}_z^{-1}(p)-R_z^{-1}(p))$ converge in distribution to mean 0 normal random variables with variances

$$\sigma_z^2(v) = \left(\frac{\partial R_z(v)}{\partial \beta}\right)\Sigma(\beta)\left(\frac{\partial R_z(v)}{\partial \beta}\right)^T + \mathrm{var}(\Psi_{1z}), \quad \text{and}$$

$$\tau_z^2(p) = \left(\frac{\partial R_z^{-1}(p)}{\partial p}\right)^2 \sigma_z^2(v),$$

for $v = R_z^{-1}(p)$ respectively. These expressions reduce to those of Theorems 1 and 2 when no covariates are modeled.

Table 1

Results of 5,000 simulations to evaluate the application of inference based on asymptotic theory to finite sample studies. Box-Cox risk with $\beta_0 = -0.486, \beta_1 = 0.793, \beta_2 = 0.4$, the same as in Table 2. Shown are results for $\hat{R}(v)$

	$v = 0.1$	$v = 0.3$	$v = 0.5$	$v = 0.7$	$v = 0.9$
$R(v)$	0.100	0.194	0.313	0.491	0.800
Bias					
% bias in $\hat{R}(v)$					
$n = 100$	−1.636	−3.058	−0.969	−0.493	−0.749
$n = 500$	−0.164	−0.619	−0.468	−0.308	−0.031
$n = 2000$	−0.279	−0.240	−0.152	−0.095	−0.033
Variance					
$\frac{\text{Asymptotic} - \text{Observed}}{\text{Observed}}\%$					
$n = 100$	−7.660	−11.976	−4.646	−7.322	7.262
$n = 500$	3.276	1.103	−0.875	−1.694	−3.330
$n = 2000$	0.538	−0.200	1.739	0.047	5.910
95% Confidence Interval					
coverage (%)					
$n = 100$	86.529	92.094	92.914	92.874	89.992
$n = 500$	92.984	94.769	94.567	94.386	94.245
$n = 2000$	94.238	95.038	95.428	95.161	95.346

Table 2

Results of simulations to evaluate the application of inference based on asymptotic theory to finite sample studies. Box-Cox risk with $\beta_0 = -0.486$, $\beta_1 = 0.793$, $\beta_2 = 0.4$, the same as in Table 1. Shown are results for $\hat{R}^{-1}(p)$

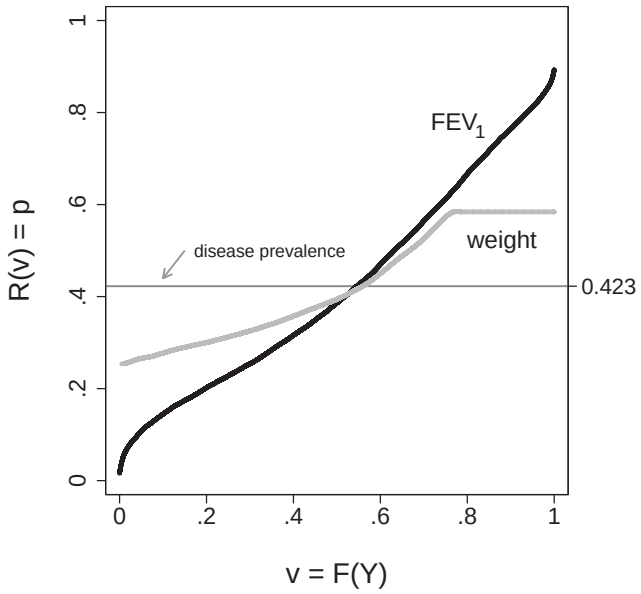
	$p = 0.1$	$p = 0.3$	$p = 0.5$	$p = 0.7$	$p = 0.9$
$R^{-1}(p)$	0.100	0.480	0.708	0.848	0.944
Bias					
% bias in $\hat{R}^{-1}(p)$					
$n = 100$	20.292	0.821	−0.892	−0.800	−0.746
$n = 500$	0.795	0.500	−0.096	−0.156	−0.199
$n = 2000$	0.316	0.154	0.0008	−0.034	−0.031
Variance					
$\frac{\text{Asymptotic} - \text{Observed}}{\text{Observed}}\%$					
$n = 100$	37.008	7.548	5.396	−9.196	−0.409
$n = 500$	11.167	1.173	−0.018	−1.133	3.857
$n = 2000$	1.249	1.730	1.134	4.121	6.095
95% Confidence Interval					
coverage (%)					
$n = 100$	73.46	92.11	92.01	94.06	91.23
$n = 500$	91.33	93.60	94.08	95.09	95.53
$n = 2000$	94.20	94.77	94.65	95.71	95.86

Table 3

Point estimates and 95% confidence intervals for $R(v)$ and $R^{-1}(p)$ using $-FEV_1$ and $-weight$ as markers of risk for subsequent pulmonary exacerbation in patients with cystic fibrosis.

	Estimate	**Variance due to $\hat{F}$**	**Confidence Interval Asymptotic**	**Confidence Interval Bootstrap**
$\boldsymbol{R(0.9)}$				
FEV_1	0.76	10.0%	(0.748,0.779)	(0.749,0.779)
weight	0.58	0.60%	(0.568,0.601)	(0.567,0.601)
$\boldsymbol{R(0.1)}$				
FEV_1	0.15	8.91%	(0.133,0.157)	(0.133,0.157)
weight	0.28	0.72%	(0.262,0.293)	(0.261,0.294)
$\boldsymbol{R^{-1}(0.25)}$				
FEV_1	0.29	18.16%	(0.273,0.311)	(0.270,0.314)
weight	0	0%	(0,0.040)	(0,0.039)
$\boldsymbol{R^{-1}(0.75)}$				
FEV_1	0.88	13.58%	(0.865,0.897)	(0.864,0.898)
weight	1	0%	(1,1)	(1,1)

Figure 1. Predictiveness curves for two markers of pulmonary exacerbation in children with cystic fibrosis. F is the cdf of the marker. The x-axis concerns the marker quantile and the y-axis shows the corresponding risk, $R(v) = P[D = 1|Y = F^{-1}(v)]$



.

Figure 2. Predictiveness curves for lung function as a predictor of risk of pulmonary exacerbation in the subsequent year. Shown are curves for cystic fibrosis children with poor, average and good nutritional status defined by weight at the 1^{st}, 2^{nd} and 3^{rd} quartiles, the incidences of pulmonary exacerbation in the 3 groups being 0.58, 0.40 and 0.30, respectively.

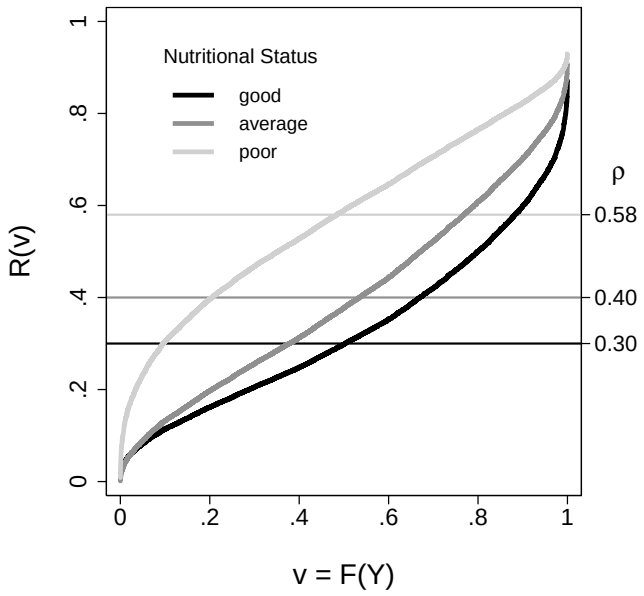


Figure 3. Predictiveness of PSA and PSA velocity as markers for prostate cancer risk. The top panel is for all cancers while the bottom panel is for the subset of high grade cancers (Gleason score > 6). The horizontal lines show the prevalence of cancer and correspond to predictiveness curves of a useless marker.