2013

The predictive hazard ratio for biomarker evaluation studies

Debashis Ghosh, *Penn State University*

The predictive hazard ratio for biomarker evaluation studies

Debashis Ghosh

Departments of Statistics and Public Health Sciences

The Pennsylvania State University

514A Wartik Building

University Park, PA,16802 U.S.A.

ghoshd@psu.edu

Summary

There is tremendous scientific and medical interest in the use of biomarkers to better facilitate medical decision making. In this article, we present a simple framework for assessing the predictive ability of a biomarker. The methodology requires use of techniques from a subfield of survival analysis termed semicompeting risks; results are presented to make the article self-contained. A crucial parameter for evaluating is the predictive hazard ratio, which is different from the usual hazard ratio from Cox regression models for right-censored data. This quantity will be defined; its estimation, inference and adjustment for covariates will be discussed. Aspects of censoring and causal inference related to these procedures will also be described. The methodology is illustrated with an evaluation of serum albumin in terms of predicting death in patients with primary biliary cirrhosis.

Key words: Association; Causal Effect; Copula; Cross-ratio; Dependence; Diagnostics.

1. Introduction

Recently, the use of biomarkers has been strongly advocated for in clinical research (Pepe et al., 2001; Biomarkers Working Group, 2001). The promise of biomarkers is that their measurement can be used to develop better patient management procedures during the medical decision-making process. This is related to the use of biomarkers as surrogate endpoints for medical studies. Surrogate endpoints are proposed based on biological considerations within a progression model of disease. One example is CD4 count levels in AIDS; the CD4 count can potentially serve as a surrogate endpoint for death. Another example from cancer studies is using tumor shrinkage as a surrogate endpoint for survival or disease-free survival.

The viewpoint taken in this paper is that the biomarker will be used clinically to trigger further decisions in practice. As an example, consider prostate cancer. Typically, prostate-specific antigen (PSA) has been used for detection of prostate cancer. If a man has a PSA measurement between 4 and 10 ng/mL, then this leads to a prostate needle biopsy. If the biopsy is positive for prostate cancer, the patient either undergoes surgical removal of the prostate (radical prostatectomy) or is monitored periodically for elevations in PSA (watchful waiting). Thus, the PSA measurement being thresholded at 4 ng/mL is often used to trigger a medical intervention. While PSA is known for being a relatively sensitive biomarker, it is not known as being a very specific measurement. As a result, many biopsies yield negative results for tumor, even when the PSA is between 4-10 ng/mL.

In many applications, associations between biomarker and a outcome is typically assessed through regression models. Here and in the sequel, we will focus on the response variable being a time to event that is potentially subject to right censoring. One commonly used model is the proportional hazards model (Cox, 1972; Gail et al., 1981). There are many studies in which univariate and possibly multivariate PH regression models are fit in which one of the covariates is the biomarker of interest.

The approach we take in this paper is to treat the time at which biomarker positivity occurs as a time to event such as that commonly used in the area of survival analysis. We then wish to study the association between the time to biomarker positivity and the time to the clinical endpoint of

interest. We formulate the observed data as data structure that we term semi-competing risks data (Fine et al., 2001; Ghosh, 2006, 2009; Ghosh et al., 2012). Based on this framework, we define a quantity known as the predictive hazard ratio. This quantity was initially proposed by Bryant et al. (1997). They argued eloquently for its use in assessing the dependence between time to a landmark event with time to a clinical endpoint. The predictive hazards ratio is quite compatible with the semi-competing risks data structure. One of its crucial features is that the region of interest is when the time to biomarker positivity occurs before the important clinical event. From our view, use of the predictive hazard ratio has several appealing features:

1. There exists a well-established theory and asymptotic results to provide inference about the predictive hazard ratio;

2. The predictive hazard ratio is very flexible in adjustment for covariates.

The structure of this paper is as follows. In Section 2, we review the data structures as well as semi-competing risks data. The proposed methodology is described in Section 3. We illustrate the application of the methodology with application to data from a primary biliary cirrhosis study in Section 4. Finally, the paper concludes with discussion in Section 5.

2. Preliminaries and Background

2.1. Data Structures

We start by making the following definitions. Let $a \wedge b$ denote the minimum of two numbers a and b. Define $I(A)$ to be the indicator function for the event A. Let T be the time to a clinical event and C time to independent censoring. Let $\mathbf{Z}(t)$ be the longitudinal process for the biomarker. We observe the data $\{Y_i, \delta_i^Y, (\mathbf{Z}_i(t); t \leq Y_i)\}$, $i = 1, \ldots, n$, n independent and identically distributed observations from $\{Y, \delta^Y, \mathbf{Z}_i(\cdot)\}$, where $Y = T \wedge C$ and $\delta^Y = I(T \leq C)$.

A standard analysis that is done here is to typically model the event time using a time-dependent regression model, such as the Cox proportional hazards with time-dependent covariates (Gail et al., 1981):

$$\lambda(t|\mathbf{Z}(s) : 0 \leq s \leq t) = \lambda_0(t) \exp\{\beta' \mathbf{Z}(t)\},$$

where $\lambda_0(t)$ denotes the baseline hazard function, and β represents the parametric regression coefficient. Estimation in such a model can be done using the partial likelihood, which in standard in many software packages. While there exist asymptotic results to guide the user in terms of estimation and inference for the estimates of the parameters in the Cox model, we can argue as in Pepe et al. (2006) that measures of association in a survival regression do not convey appropriate information about biomarker utility in terms of classification. This necessitates looking at other measures to evaluate biomarkers, such as what is being proposed in this article.

2.2. Semi-competing risks

For the proposed framework in the paper, we simply convert the covariate process $\mathbf{Z}(t)$ into another failure time variable X (possibly censored) based on a positivity criterion. Here, the observed data become $(X_i, \delta_i^X, Y_i, \delta_i^Y)$, $i = 1, \ldots, n$, n independent and identically distributed observations from $(X, \delta^X, Y, \delta^Y)$, where $X = S \wedge T \wedge C$, $\delta^X = I(S \leq T \wedge C)$, $Y = T \wedge C$ and $\delta^Y = I(T \leq C)$.

The proposed framework defines a new random variable S that is the time to biomarker positivity. Note that we have censored S, the time of biomarker positivity, by the minimum of T and C and not just by C. From the biomarker point of view, it is of no use if S is larger than T. In this instance, the biomarker becoming positive will occur after the clinically relevant event does, so its being positive provides no practical utility in aiding the management of the patient. In addition, from a practical point of view, the biomarker becoming positive will trigger some type of medical intervention. This suggests that the region of practical interest for the joint distribution of (S, T) occurs when $S < T$. We will refer to this region as the wedge. Fine et al. (2001) prove that it is possible to identify the joint distribution on the wedge. This type of data is called *semi-competing risks data* because of the inherent asymmetry in the dependence and censoring structures for S and T.

To summarize the effect of the biomarker, we will use what Bryant et al. (1997) term the predictive hazard ratio. The model being assumed is that of Clayton (1978) and Oakes (1982) and can be formulated as the following:

$$\theta = \frac{\lambda_T(t|S = s)}{\lambda_T(t|S \geq s)}, \tag{1}$$

where $\lambda_T(t|A) = \lim_{\Delta t \to 0} \frac{d}{dt} Pr(T < t + \Delta t | T \geq t, S \in A)$, and A is a subset of the interval $(0, \infty)$. The right-hand side of (1) is the predictive hazard ratio and depends on both s and t. However, the left-hand side does not.

Note that the proposed model is being formulated only for the wedge. There will be uncountably infinite joint distributions for (S, T) that are consistent with the model (1). One trivial example is to assume the model for the entire joint distribution of (S, T). We discuss possible conceptual problems with this approach within a causal framework in §3.3.

To estimate θ, we follow the approach of Fine et al. (2001). They provided a closed form estimator of θ using modified weighted concordance estimating functions from Oakes (1982, 1986) along with an asymptotic variance estimator. For $i = 1, \ldots, n$ and $j = 1, \ldots, n$, define $\tilde{X}_{ij} = X_i \wedge X_j$, $\tilde{Y}_{ij} = Y_i \wedge Y_j$, $\tilde{C}_{ij} = C_i \wedge C_j$ and $D_{ij} = I(X_{ij} < Y_{ij} < C_{ij})$. Fine et al. (2001) proposed the following closed-form estimator for θ:

$$\hat{\theta} = \frac{\sum_{i<j} W(\tilde{X}_{ij}, \tilde{Y}_{ij}) D_{ij} \Delta_{ij}}{\sum_{i<j} W(\tilde{X}_{ij}, \tilde{Y}_{ij}) D_{ij} (1 - \Delta_{ij})},$$

where $W(u, v)$ is a weight function that converges uniformly to $w(u, v)$, a bounded deterministic function, and $\Delta_{ij} = I\{(X_i - X_j)(Y_i - Y_j) > 0\}$, $i, j = 1, \ldots, n$. They also prove the consistency and asymptotic normality of $\hat{\theta}$ using a combination of U-statistic theory. Here and in the sequel, we will take the weight function to be unity.

The variances for the limiting distribution of these random variables are fairly complicated. Here, we will use a resampling method proposed by Jin, Ying and Wei (2001) and used in Ghosh (2009). The algorithm proceeds as follows:

1. We generate n Exponential random variables $(G_1, \ldots, G_n)$ and calculate $\hat{\theta}^*$, where

$$\hat{\theta}^* = \frac{\sum_{i<j} D_{ij} \Delta_{ij} G_i G_j}{\sum_{i<j} D_{ij} (1 - \Delta_{ij}) G_i G_j},$$

2. Repeat step 1 M times.

3. Estimate the variance of $\hat{\theta}$ based on the empirical variance of $\hat{\theta}^*$.

This resampling procedure is quite fast. In practice, we usually take $M = 1000$. It is very similar in concept to the bootstrap (Efron and Tibshirani, 1986). In terms of theoretical justification, it

can be proven using arguments as in Ghosh (2009) that the conditional distribution of $\hat{\theta}^* - \hat{\theta}$ given the detail is asymptotically the same as the unconditional distribution of $\hat{\theta} - \theta$. This result formally justifies the use of the resampling algorithm described above.

3. Proposed Methodology

3.1 Biomarker Evauation

As alluded to in Section 2, we convert the problem of biomarker evaluation into one of estimating the predictive hazard ratio. We will assume throughout the paper that positivity occurs when the biomarker reaches above a cutoff value c. Define S as the time to this event. We then create the bivariate survival dataset as described in Section 2.2 and calculate the estimate of $\theta \equiv \theta(c)$ as well as the associated 95% confidence intervals. We then vary the values of c and plot the estimates of the dependence parameter and the associated 95% CI. This then provides a useful method of presenting results on the discriminative ability of the biomarker. In particular, one might consider regions of the plot in which $\hat{\theta}$ is highest. If there is no such region, then this suggests that the choice of biomarker cutoff has minimal effect on its predictive power. A related inferential problem is that one might wish to see if the biomarker predicts better than 'random chance.' Within our framework, this corresponds to testing $H_0 : \theta = 1$ and can be done by checking if the associated confidence interval contains 1 or not.

It should also be noted that the event time S can be defined in many ways. One way is to define it as the time to a biomarker above a cutoff. However, it could also be defined based on multiple biomarker measurements or a statistic thereof. For example, one could calculate some type of moving average and define S to be the time at which the moving average is above a certain cutoff. Alternatively, one could calculate a slope based on a moving window of measurements and determine S based on time to when the slope is above a cutoff. The proposed framework is quite flexible in how the biomarker gets utilized to calculate the value of S.

3.2 Covariate Adjustment

In many scientific settings, the distribution of the biomarker will depend on other covariates, such as gender, race and other confounding factors. Thus, it is important to be able to adjust for

covariates in the analysis.

The issue of covariate adjustment becomes quite complex in the current modelling framework. There are many possible ways in which covariates can affect the joint distribution of (S, T). First, the distribution of the biomarker can depend on other variables. Examples of such variables would include age and gender. In this scenario, one would formulate models for the distribution of $\mathbf{Z}(t)$ conditional on covariates; examples of such models include linear mixed-effects models (Laird and Ware, 1982). Based on the fitted values from such a model, we then define S based on a covariate-adjusted biomarker positivity criterion and apply the methods as described before. Such an approach is quite straightforward to implement.

One concern then becomes that the variability in the covariate-adjusted dependence parameter estimate comes from two sources: (a) the variance in the estimate of the biomarker covariate adjustment model; (b) the variance in the dependence parameter estimate. There are two modes of inference we can employ. The first is termed conditional and ignores step (a) in the variability estimation. For this approach the standard error calculation described in §2.2. is sufficient; this is what is used in the examples presented in §4. The other mode of inference is unconditional and attempts to incorporate both sources of variability from (a) and (b). We can use the nonparametric bootstrap, repeat the two-stage estimation process, and use the bootstrapped empirical distribution of the dependence parameter estimators to calculate variance estimates and construct confidence intervals. Such an approach will lead to wider confidence intervals relative to those shown in Figures 2 and 3.

An alternative method of covariate adjustment is to assume that the predictive hazard ratio depends on covariates. If one assumes that the variables being considered are effect modifiers of the predictive hazard ratio so that the association between biomarker positivity and the event of interest depends on the combination of variables, then one could compute stratum-specific predictive hazard ratio estimators, where the strata are defined by the combination of covariate levels.

3.3 Causal Inference implications

There has been a lot of interest in assessing causal effects of biomarkers within a surrogate endpoints framework (Taylor et al., 2005; Gilbert and Hudgens, 2008; Joffe and Greene, 2009). A

natural question which can then be asked is whether or not the proposed framework here enjoys a causal inference interpretation. This issue was addressed recently in Ghosh (2012). There, several results were found. First, while much of the biostatistical literature formulates causal effects in terms of the potential outcomes model (Rubin, 1974; Holland, 1986), such a framework in incompatible with the type of dependent censoring considered. The reason is that intuitively, the wedge restriction makes it impossible to define potential outcomes for (S, T). One can then resort to a different causal modelling framework that has been adopted in econometrics (Abbring and van den Berg, 2003) to determine an equivalence with semi-competing risks data. Since this model is different from the potential outcomes framework, we briefly describe it here. Conditional on a nonnegative bivariate random vector $\mathbf{V}_Z = (V_{SZ}, V_{TZ})$

$$\lambda_{SZ}(s|\mathbf{V}_Z) = V_{SZ}\lambda_{0S}(s|Z) \tag{2}$$

$$\lambda_{TZ}(t|\mathbf{V}_Z) = \begin{cases} V_{TZ}\lambda_{0T}(t|Z) & \text{if } t < S \\ V_{TZ}\lambda_{0T}(t|Z)\theta(t|S,Z) & \text{if } t \geq S \end{cases} \tag{3}$$

where λ_{0S} and λ_{0T} denote hazard functions for S and T, conditional on covariates, and $\theta(t|S, Z)$ denotes the effect of S on T (conditional on Z) after the event occurs. The vector $\mathbf{V}_Z$ is known as the frailty and is used to account for selection effects in the population (e.g., Hougaard, 1986). Conceptually, model (2)-(3) operationalizes S as a type of "intervention" that is applied to the individual. After the intervention occurs, the effect is to modify the hazard by the function $\theta(t|S, Z)$. Note that this type of model is more compatible with the structural models considered in Pearl (2009) rather than the potential outcomes framework. While Ghosh (2012) showed that the model (2-3) was compatible with semicompeting risks data, he also demonstrated that for this framework, strong predictive hazard ratios did not necessarily imply strong causal effects of biomarker positivity. As described there, attempting causal inference in these settings will typically require strong mechanistic knowledge about the pathways through which the biomarker is affecting the true endpoint.

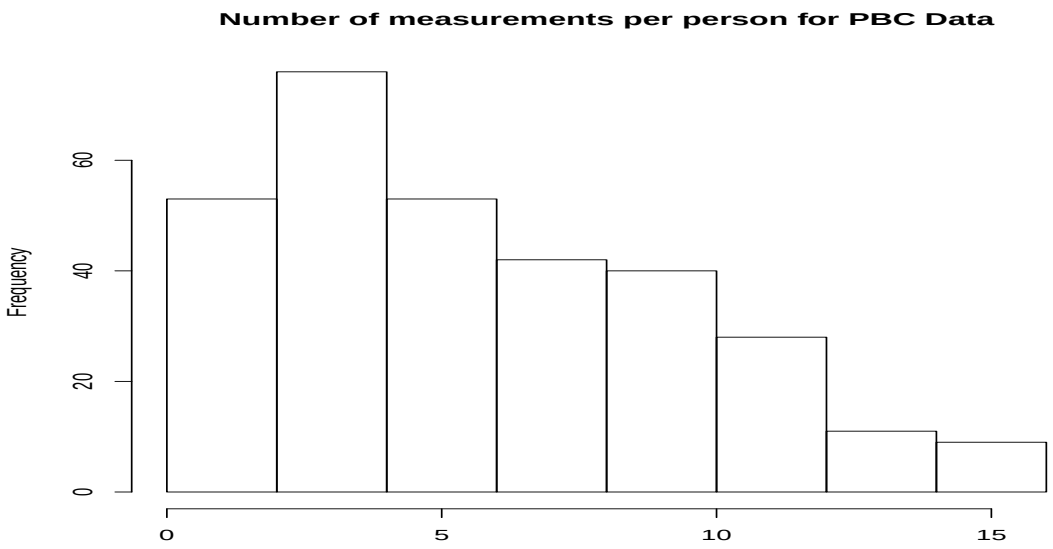
3.4 Censoring Mechanisms

It should be pointed out that the biomarker positivity time is assumed to be right-censored for the methods in this paper. This might not always be possible in practice. For example, let

us consider shrinkage of a colorectal cancer tumor 6 months as the biomarker; the relevant time to event is overall survival. The percentage of shrinkage is known. Biomarker positivity then can be based on whether or not the tumor shrinks by a certain percentage p_0. Using the notation of the paper here, one could define S to be the time until the tumor shrinks by p_0 and T to be time to death. However, S then becomes considered as current status data, because the tumor measurement is made only at six months. Thus, the information available about S here is whether or not it occurred before six months. Development of semi-competing risks procedures in such a framework is quite challenging and beyond the scope of the current article.

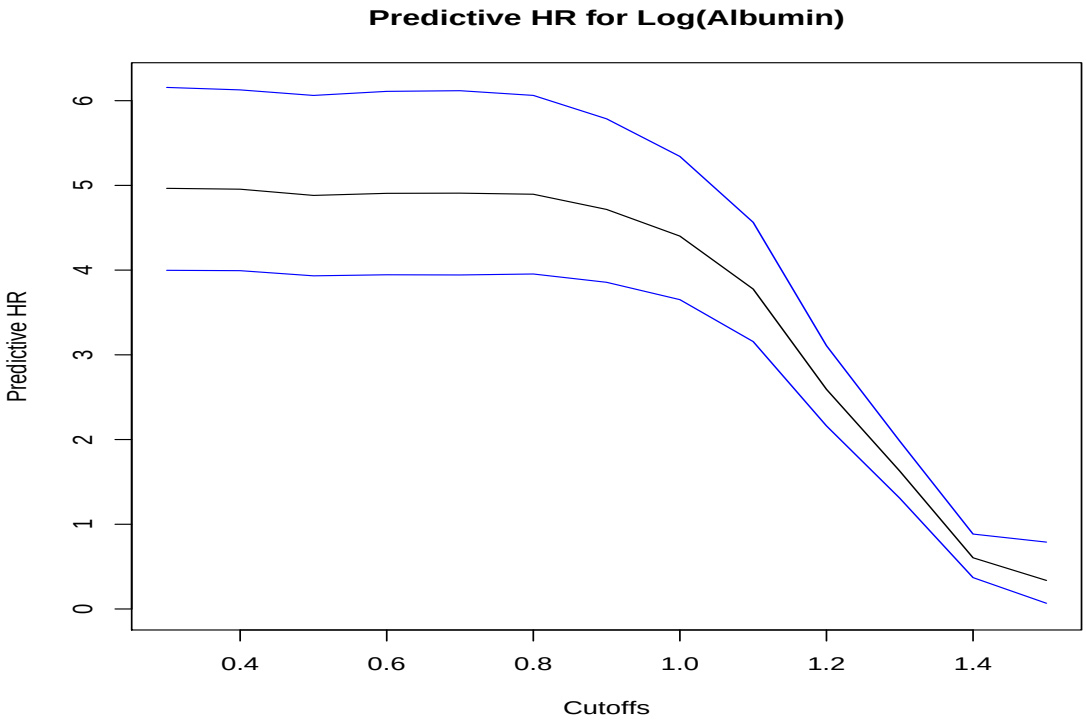
4. Numerical Example: PBC data

The data we consider in this paper are from a famous primary biliary cirrhosis study that is available as an appendix in Fleming and Harrington (1991). We work with an extended version of the dataset that was analyzed by Murtaugh et al. (1994). These data involve repeated visits by the patients who were diagnosed with PBC and seen at the Mayo Clinic between January 1974 and May 1984. In particular, there are 312 subjects generating a total of 1945 measurements. The distribution of visits per person can be found in Figure 1. In Murtaugh et al. (1994), the goal of the study was twofold. One was to update the Mayo model for predicting survival in subjects with PBC using repeated measurements. The model consists of the following variables: age, bilirubin, prothrombin time, albumin and edema. To illustrate the procedures developed here, we will focus on bilirubin, which is a biochemical measurement indicative of liver activity.



As in prior analyses of these data, we will transform the albumin measurements to a log scale in order to reduce skewness. If we fit a proportional hazards model using only the albumin measurement at baseline, then the log hazard ratio is 5.4, with an associated standard error of 0.6. This yields a statistic that is strongly associated with risk of death (p-value $< 2 \times 10^{-16}$). Next, we consider a time-dependent PH model with log(albumin) as the covariate. It turns out that the estimated log hazard ratio does not change much from before, but the standard error is reduced to 0.33. As suggested by the arguments of Pepe et al. (2006), such a strong association does not necessarily mean that albumin is useful for prediction purposes.

The results of the predictive hazard ratio along with the associated 95% pointwise CIs are provided in Figure 2 below.

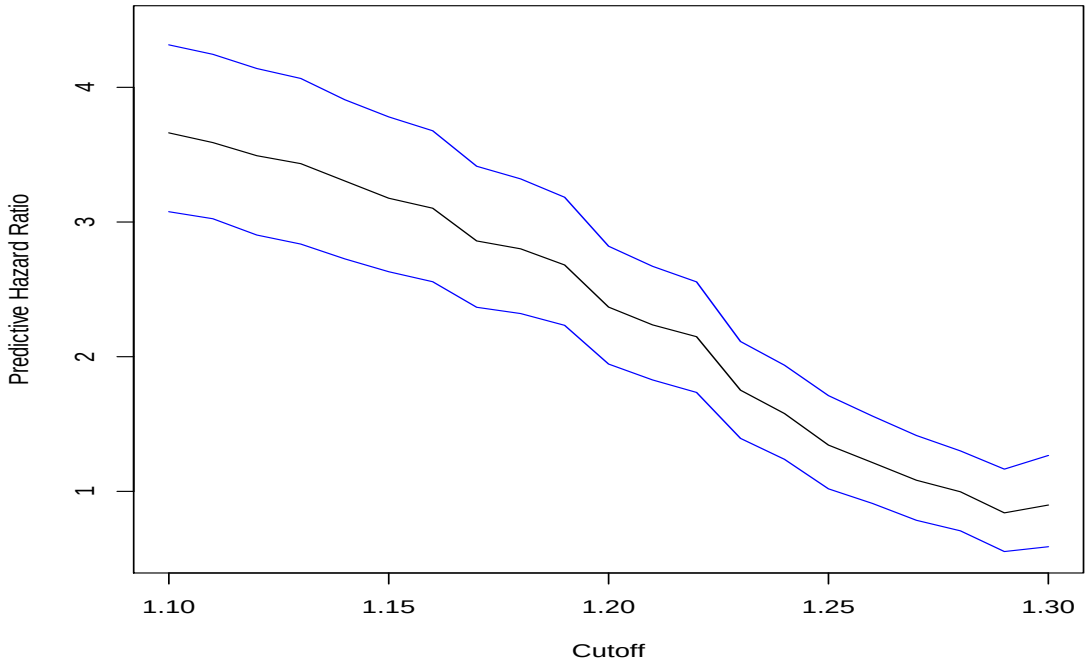


Based on the plot, we find that there is a noticeable decrease in the predictive hazard ratio corresponding a cutoff of one for albumin on the natural logartihm scale. Also, we see that the predictive hazard ratio is smaller for all cutoff values considered relative to the regression coefficients from the proportional hazards models that we previously fit. This confirms the argument by Pepe et al. (2006) that strong regression coefficients do not immediately imply strong prediction performance.

Next, we consider adjusting for covariates. We refer back to the Mayo model studied in Murtagh et al. (1994) and consider adjustment of albumin for other covariates from that model. Next, we fit a linear model regressing albumin on a logarithmic scale as a function of age at baseline, gender, edema, prothrombin time and bilirubin. The latter two covariates are log-transformed. This leads to the following regression equation for albumin:

$$\log(\widehat{\text{Albumin}}) = 2.01 - 0.001\text{Age} - 0.03\text{Female} - 0.276\log(\text{Protime}) - 0.11\text{Edema} - 0.049\log(\text{Bili}).$$

If we adjust albumin using this regression equation and compute the predictive hazard ratio, the results are presented in Figure 3. The shape of the curve is different from that for unadjusted albumin in Figure 2. There is no cutoff at which the curve has a steep decrease relative to that in Figure 2.



We again see that the predictive hazard ratio estimates are smaller than from the Cox proportional hazards model that was initially fit. This suggests less strength of evidence for the use of albumin in a medical decision making context than what is suggesting by the proportional hazards model.

5. Discussion

For the evaluation of biomarkers, it is necessary to consider prediction performance. In this article, we have presented a simple approach to prediction assesment of biomarker rules using dependent censoring methodology from survival analysis. The methodology is quite easy to implement, and code implementing the predictive hazard ratio estimation can be downloaded as supplementary material. One appealing feature is that the proposed methodology focuses on predictive ability rather than pure association; this is done through considering the wedge region that is common in much of the literature on semicompeting risks data.

There are many open problems that need to be investigated further. One is to consider biomarker positivity under more complex censoring schemes, such as the interval censoring scheme described in §3.3. Another issue is incorporating multiple biomarkers. A simple approach would be to develop a risk score that is a one-dimensional summary of the measurements and to then apply

the methodology in this paper. However, there may exist more powerful methods of modelling multiple biomarkers.

Finally, in many settings, longitudinal serum or tissue samples have been collected prospectively in many studies. This allows for the use of nested case-control designs for assessing the discriminative ability of biomarkers (Baker et al., 2002). Conceptually, this involves following subjects prospectively in order to determine disease status; based on the disease status, the distribution of biomarkers between cases and controls are then compared. It would be of interest to extend the predictive hazard ratio to this setting as well.

There is code available for implementing the methods in this paper and for the example in Section 4. It is available at `www.bepress.com/debashis_ghosh/`.

Acknowledgments

The project described was supported by NIH R01-CA129102 and the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR000127. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Abbring, J. H. and van den Berg, G. J. (2003). The nonparametric identification of the treatment effects in duration models. *Econometrica* **71**, 1491 – 1517.

Baker, S. G., Kramer, B. S. and Srivastava, S. (2002). Markers for early detection of cancer. Statistical guidelines for nested case control studies. *BMC Med. Res. Methodol.* **2**, 4.

Biomarkers Working Group. (2001). Biomarkers and surrogate endponts: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* **69**, 89 – 95.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141 – 151.

Day, R., Bryant, J. and Lefkopolou, M. (1997). Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika* **84**, 45 – 56.

Efron, B. and Tibshirani, R. (1986). Bootstrap method for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science* **1**, 54 – 77.

Fine, J. P., Jiang, H. and Chappell, R. (2001). On semi-competing risks data. *Biometrika* **88**, 907 – 919.

Fleming T and Harrington D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Gail, M. (1981). Evaluating serial cancer marker studies in patients at risk of recurrent disease. *Biometrics* **37**, 67 – 78.

Ghosh, D. (2006). Semiparametric inferences for association with semi-competing risks data. *Statistics in Medicine* **25**, 2059 – 2070.

Ghosh, D. (2009). On assessing surrogacy in a single-trial setting using a semi-competing risks paradigm. *Biometrics* **65**, 521 – 529.

Ghosh, D. (2012). A causal framework for surrogate endpoints with semi-competing risks data, *Statistics and Probability Letters* **82**, 1898 – 1902.

Ghosh, D., Taylor, J. M. and Sargent, D. J. (2012). Meta-analysis for surrogacy: accelerated failure time modelling and semi-competing risks (with discussion), *Biometrics* **68**, 226 – 247.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64**, 1146 – 1154.

Holland, P. 1986. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81**, 945 – 970.

Jin, Z., Ying, Z. and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381 – 390.

Joffe, M. M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65**, 530 – 538.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963 – 974.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L., and Gips, C. H. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology* **20**, 126-34.

Oakes, D. (1982). A model for association in bivariate survival data. *J. R. Statist. Soc.* B **44**, 414 – 422.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73** 353 – 361.

Pearl J. (2001). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

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

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2006). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* **159**, 882 – 890.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688 – 701.

Taylor, J. M. G, Wang, Y. and Thiebaut, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**: 1102-1111.