Statistical Methods for Analyzing Tissue Microarray Data

Xueli Liu[1, 2], Vladimir Minin[3], Yunda Huang[1], David B. Seligson[4], and Steve Horvath[1, 2, *]

[1] Department of Biostatistics, School of Public Health, UCLA

[2] Department of Human Genetics, David Geffen School of Medicine, UCLA

[3] Department of Biomathematics, David Geffen School of Medicine, UCLA

[4] Department of Pathology, David Geffen School of Medicine, UCLA

[*] Correspondence: Steve Horvath, Department of Human Genetics, Gonda Research Center, David Geffen School of Medicine, UCLA, 695 Charles E. Young Drive South, Box 708822, Los Angeles, CA 90095-7088, USA; Fax: 810-277-7453; E-mail: shorvath@mednet.ucla.edu.

ABSTRACT

Tissue microarrays (TMAs) are a new high-throughput tool for the study of protein expression patterns in tissues and are increasingly used to evaluate the diagnostic and prognostic importance of biomarkers. TMA data are rather challenging to analyze. Covariates are highly skewed, non-normal, and may be highly correlated. We present statistical methods for relating TMA data to censored time-to-event data. We review methods for evaluating the predictive power of Cox regression models and show how to test whether biomarker data contain predictive information above and beyond standard pathology covariates. We use nonparametric bootstrap methods to validate model fitting indices such as the concordance index. We also present data mining methods for characterizing high risk patients with simple biomarker rules. Since researchers in the TMA community routinely dichotomize biomarker expression values, survival trees are a natural choice. We also use bump hunting (patient rule induction method), which we adapt to the use with survival data. The proposed methods are applied to a kidney cancer tissue microarray data set.

Key Words: Survival tree; Bump hunting; C-index; Cox regression; Model validation

1 INTRODUCTION

Tissue microarrays (TMAs; Kononen et al., 1998) make high throughput molecular analysis of large numbers of tumor samples in a single immunohistochemical staining reaction possible. TMAs are a tool to validate the role of newly-identified tumor biomarkers. With additional molecular information and appropriate statistical models, biomarkers are expected to lead to improved diagnostic, prognostic and therapeutic applications in the clinic. For example, the cell proliferation biomarker Ki-67, has been shown to be significantly associated with survival in prostate cancer (Bettencourt et al., 1996).

The basic tissue array technique is summarized in Kononen et al. (1998). Hundreds of tiny (typically 0.6 mm diameter) cylindrical tissue cores are densely and precisely arrayed into a single histologic paraffin block. The block may be divided into up to 300 serial 4 to 8 μm thick sections, which we refer to as tissue array slides (figure 1). These tissue array slides serve as targets for immunohistochemical staining reactions. Each tissue array slide yields information about protein staining pattern, distribution, intensity, background, and target tissue. TMAs are widely used in determining cellular expression values and tissue distribution patterns for newly-identified genes on a variety of normal and diseased tissue specimens.

In this paper, we propose statistical methods for relating TMA data to right censored failure times, e.g. post-operative survival or time to first tumor recurrence. Censored times are an important outcome in practice, but it is straightforward to adapt our methods to uncensored or binary outcomes.

The paper is organized as follows: section 2 describes TMA data; section 3 presents methods for showing that TMA data predict survival when biomarker expression measures are defined prior to looking at clinical outcomes; section 4 presents rule induction methods for characterizing high risk patients in terms of TMA data; section 5 shows an application to kidney cancer TMA data; section 6 discusses the results.

2 TISSUE MICROARRAY DATA

A typical TMA data set consists of three parts on different levels of observation. On the highest, most aggregated level, one deals with *clinical* patient information, e.g., age at surgery, health performance status $ECOG = 0, 1, 2, 3$, post-operative survival time, etc.

On a lower level, one deals with pathology *case* information. A case is a surgical event, from which representative tissues are taken and sampled into the tissue array. A patient may have several surgical events and so several pathology cases may correspond to one patient. For simplicity, we assume that there is a one-to-one relation between clinical patient and pathology case data represented on each tissue array. Pathology case covariates include tumor morphology information such as T-stage ($tstage = 1, 2, 3, 4$), metastasis status (met, binary) and tumor grade ($grade = 1, 2, 3, 4$).

On the lowest level, one deals with spots on a particular tissue array slide. A typical tissue array slide is presented in figure 1. Each patient is represented by multiple spots in a tissue microarray. We assume here that each tissue array slide is immunohistochemically stained by a single biomarker. For a particular spot, stained by a single biomarker, a pathologist arrives at several *staining scores* that measure biomarker expression. In our real data, three staining scores have been measured: i) the maximum staining intensity Max $\in \{0, 1, 2, 3\}$, ii) the percent of cells staining Pos $\in \{0, \ldots, 100\}$, and iii) the percent of cells staining with maximum intensity PosMax $\in \{0, \ldots, 100\}$. For a given marker, the staining scores (Max, Pos, and PosMax) tend to be highly correlated. The distribution of Pos tends to be highly skewed (figure 2) or semi-continuous. For example, the Pos value of the majority of spots may be 0 but follow a continuous distribution for the remaining spots. To denote which staining score is measured for a particular biomarker, we usually append the staining score method to the biomarker name, e.g., P53Max, P53Pos, or P53PosMax.

In this paper, we propose methods for relating staining scores of biomarkers to clinical outcome information. Thus, there is a need to pool multiple spot measurements across each case (patient). TMA *case* data contain *pooled* estimates of spot biomarker staining scores across each case (patient). The simplest pooling methods are to form the mean, median, maximum or minimum value of the spot measurements. We denote these corresponding pooled measurements by appending .mn, .md, .max, and .min, respectively to biomarker variable names. For example, P53Max.mn denotes the mean pooled maximum intensity staining score of biomarker P53.

As is to be expected, the different spot pooling methods result in highly correlated staining scores, e.g., Pos.mean will usually be highly correlated with Pos.md. Thus, one arrives at multiple highly correlated pooled staining scores: Pos.mn, Pos.md, . . ., Max.mn, Max.md, etc. To add to this multiplicity, consider that it is a standard practice in the TMA community to dichotomize pooled staining scores since over- and under-expression of a biomarker lends itself to easy biological interpretation. For each biomarker, one arrives at numerous highly correlated dichotomous biomarker expression scores. Explicitly, the number is the product of the number of staining scores, the number of pooling methods, and the number of potential cut-off values. We have found that different choices can lead to covariates with very different significance levels in subsequent Cox regression models (Cox, 1972).

2.1 Defininining Biomarker Expression Indices

A challenge is to pick a staining score (e.g. Max, Pos, or PosMax), a pooling method, and possibly a cut-off value for each biomarker. Ideally, these choices should be guided by prior biological knowledge. For example, the cell proliferation marker Ki-67 is usually measured by its Pos score. The pooling method should also reflect knowledge about the biological action of the protein. For example, mean pooling would be appropriate if one expects that the average (global) biomarker expression is related to survival. When these choices are made prior to looking at the outcome data, it is straightforward to test whether the biomarker is significantly associated with survival time (see the section on using TMA data to predict survival).

However, if no prior biological knowledge is available on how to score, pool, and dichotomize a biomarker, it is of interest to develop rules for constructing biomarker covariates. For this case, we present several data mining methods for evaluating how to dichotomize and combine biomarker staining scores (see our section on characterizing high risk patients with biomarker rules).

3 USING TMA DATA TO PREDICT SURVIVAL

Here we study predictive models of patient survival when the biomarker covariates were defined prior to looking at the clinical or pathology data. The standard method is to use the biomarker covariates along with clinical covariates in a Cox regression model (Cox, 1972). The Cox regression model assumes non-informative censoring and proportional hazards, which can be verified with the use of Schoenfeld residuals (Schoenfeld, 1982).

Below we will construct three different Cox models: a clinical model with clinical and pathology covariates only, a biomarker model with biomarker covariates only, and a biomarker/clinical model based on both biomarker and clinical covariates. An important scientific question is to study whether biomarkers add predictive information above and beyond clinical and pathology covariates.

To measure the predictive accuracy (discriminatory power) of different Cox models, we use the condordance (C-) index (Harrel et al., 1982; Harrel et al., 1984; Harrel, 2001) and the

Nagelkerke's R^2 (Nagelkerke, 1991). We will briefly describe both model fitting indices below. The C-index is increasingly used in the medical literature to assess the discriminatory power of a survival model (Aaronson et al., 1997; Hachamovitch et al., 2003; Clark et al., 2003). This index is a generalization of the area under the Receiver Operating Characteristics (ROC) curve to survival outcomes and is closely related to Somers' D_{xy} rank correlation (Somers, 1962), $D_{xy} = 2(C - 0.5)$. The C-index is the proportion of all pairs of subjects whose survival time can be ordered such the subject with the higher predicted survival is the one who survived longer. The subjects' survival times cannot be ordered if both subjects are censored or if one has failed and the follow-up time of the other is less than the failure time of the first. The C-index is a probability of concordance between predicted and observed survival, with $C = 0.5$ for random predictions and $C = 1$ for a perfectly discriminating model. A C-index of 0 indicates that the 'opposite' predictor has perfect discriminatory power. Nagelkerke's R^2 index is defined as $R^2 = \frac{1-(L(0)/L(\hat{\beta}))^{2/n}}{1-L(0)^{2/n}}$ where $L(\hat{\beta})$ and $L(0)$ denote the Cox partial likelihoods of the fitted and the null (intercept only) model, respectively. Nagelkerke's R^2 ranges from 0 to 1. In our analyses, we use functions in the *Design* and *Hmisc* libraries (Harrel, 2001) of the R software (Ihaka and Gentleman, 1996).

To protect against overfitting due to including multiple covariates in a Cox model, we compute 'validated' model fitting indices to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects or subjects not used to develop our model. Specifically, the enhanced bootstrap (Efron, 1983) is used to estimate the bias due to overfitting or the 'optimism'. After the optimism is estimated, it is subtracted from the corresponding model fitting index that was derived from the original sample to obtain a bias-corrected estimate; see the *validate* of the Design library in R (Harrel, 2001).

The validated model fitting indices R^2 and C are used to rank the different models. Harrel

(2001) introduced a U-test statistic based on the C-index, which has asymptotically a standard normal distribution under the null hypothesis of no difference in predictive power (rcorrp.cens function of the Design library). This function computes the rank correlation for paired predictors with a censored response. The percentage of pairs is determined for which one model correctly selects the patient with the longer survival time while the competing model does not.

We have found that clinicians often appreciate a nomogram to visualize the final Cox regression model (figure 3). A nomogram is a visualizing aid to obtain predicted values manually from a Cox model. We will briefly review how to evaluate the nomogram. For each predictor, read the points assigned on the $0 - 100$ scale and add these points. Read the results on the 'Total Points' scale and then read the corresponding predictions below it. For example, if a patient has met$= 1$, CA9MemPos.mn$= 0$, p53Pos.mn$= 0$, pTENPos.mn$= 100$, VimPos.mn$= 0$, ECOG$= 1$, grade$= 2$, tstage$= 1$, the total number of points is $86 + 65 + 33 + 10 = 194$. The predicted 3-year survival rate is about 0.53 and the median survival time is about 3.7 years.

4 CHARACTERIZING HIGH-RISK PATIENTS

Biologists are sometimes interested in simple rules involving biomarkers that characterize high risk or low risk patients. In this data exploration phase, one is interested in rule induction methods. For example, when dealing with a particular biomarker a biologist may want to know which staining score (Max or Pos), pooling method, and cut-off value should be used to arrive at a binary biomarker covariate that optimally distinguishes high risk from low risk patients. When dealing with multiple biomarkers one may wonder how to characterize high risk patients with multiple biomarker scores. Since it is customary in the TMA community to dichotomize biomarker expression values, it is natural to use tree-predictors (Breiman et al., 1984; Zhang and Singer, 1999; Ahn and Loh, 1994) or bump hunting (Friedman and Fisher, 1999) for these questions. Both methods are described in more detail below.

Numerous methods have been proposed for fitting tree structured predictors to censored failure times (Zhang and Singer, 1999; Ahn and Loh, 1994). A simple approach is to use martingale residuals that result from fitting an intercept-only Cox regression model to the censored survival times as (uncensored) outcome in a regression tree (Therneau et al., 1990). For the ith subject the martingale residual is defined as $M_i = \delta_i - \Lambda_0(t_i)$ where t_i is the possibly censored survival time, Λ_0 is an estimate of the baseline cumulative hazard function, and δ_i is the censoring indicator. Here we will use deviance residuals (LeBlanc and Crowley, 1992), which have a more symmetric distribution than martingale residuals. For the ith subject the deviance residual is defined as

$$d_i = sign(M_i)\sqrt{2[\delta_i log(\frac{\delta_i}{\Lambda_0(t_i)}) - M_i]}. \tag{1}$$

LeBlanc and Crowley (1992) demonstrated a) that using deviance residuals in regression trees is similar to the survival tree methods presented by Segal (1988) and Ciampi et al. (1986), and b) that using deviance residuals is more efficient than using martingale residuals with regression trees. To fit regression trees to the deviance residuals, we use the default settings of the rpart function in R: the nodes are split with the Anova method, which is equivalent to maximizing the between-groups sum-of-squares. Future studies should investigate how this compares to alternative survival tree methods, e.g. to (Ahn and Loh, 1994)

Compared to the Cox regression model, survival trees have several advantages. First, there are no problems with convergence when dealing with multiple highly correlated covariates. This is a primary reason for using them when analyzing the staining scores of a biomarker. Second, there is no need to validate the proportional hazards assumption. Third, the results are intuitively interpretable and easier to understand for non-statisticians. One major drawback of tree based methods is that at each step the data are recursively partitioned into two parts, which constrains the rules to follow a binary tree structure. Most tree based methods are based on a *greedy* minimization method, which may run out of data before examining important interactions.

Bump hunting was introduced by Friedman and Fisher (1999) as an alternative to tree predictors. It is not constrained to a tree structure and is a more *patient* rule induction method in the sense that it has a more efficient way of learning from the data. Bump hunting allows one to recover complex interactions between covariates together with their cut-off values. The goal of bump hunting is to partition the feature (covariate) space into box-shaped regions seeking boxes with a high average of the response variable. Here we propose to use the deviance residual as surrogate for the censored survival time. The algorithm starts with a box containing all the data, and proceeds with top-down peeling until the box contains a user-defined minimal proportion of the data. At each peeling step the box is compressed along one face such that the proportion α of observations is peeled off and the removed box b^* satisfies:

$$b^* = \arg \max_{b \in C(b)} ave[y_i | \overline{x}_i \in B - b]$$

Here $\overline{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})$ - predictors, y_i - response, C(b) is a class of sub-boxes eligible for removal. After top-down peeling is complete, the resulting box is extended along a face if this results in an increase of the box mean (bottom-up pasting). After arriving at the final box, the data in the box are removed and the procedure is repeated on the reduced data set in order to find the next box (bump). In our studies, we use the bump hunting procedure implemented in the S-Plus function *supgem* by Friedman and Fisher (1999). This function comes with several diagnostic tools that help the user to refine the resulting rules (e.g., one can study how robust the rule is to choices of cut-off values or one can detect redundant covariates).

4.1 Biomarker Rules and Kaplan Meier Curves

Rules partition the patients into several groups. It is natural to use Kaplan Meier curves (Kaplan and Meier, 1958) to visualize the corresponding survivorship functions and to use the logrank test statistic to arrive at a numeric measure of curve separation. But the logrank p-value should only be considered as a descriptive measure since the survival outcomes were used to define the rule, i.e., there is severe overfitting.

Each rule for high risk patients can be encoded in a binary covariate. This covariate should not be used in a Cox regression model involving the survival outcome used in the rule construction since this will severely overfit the data. In particular, it is not appropriate to use survival data to find optimal cut-off values for dichotomizing a staining score, and then to use the dichotomized covariate in a Cox regression model (Altman et al., 1994; Heinzl, 2000). However when using the dichotomized covariate in a univariate Cox regression model, one can correct the resulting p-value for overfitting by using a formula presented in Altman et al. (1994):

$$p_{corrected} = \phi(z)[z - 1/z]log(\frac{(1-\epsilon)^2}{\epsilon^2}) + 4\frac{\phi(z)}{z}, \tag{2}$$

where ϕ is the density function of the standard normal, z is the $(1 - p_{value}/2)$-quantile of the standard normal distribution, and ϵ is the proportion of smallest or largest values that are not considered as potential cut-off values. In our TMA data analysis, we usually choose $\epsilon = 0.1$. Schumacher et al. (1997) propose to use a bootstrap procedure to correct for the dichotomization bias in *multi-variable* Cox regression models.

5 KIDNEY CANCER TMA DATA

We used TMA data on 8 biomarkers involving 318 renal cell carcinoma cases. Methods for collection and analysis of laboratory specimens were described in Bui et al (2003). Renal cell carcinoma (RCC) is the most common cancer of the kidney. Its complex natural history cannot be completely explained by clinical prognostic factors such as grade, stage and tumor size (Pantuck et al., 2001). It is an important question whether the biomarkers contain predictive information for cancer survival. For most of the biomarkers, the staining scores Max, Pos and PosMax were measured. Then the staining scores were pooled with four different pooling methods, which resulted in 116 biomarker staining scores.

Table 1 lists the results of univariate Cox regression models involving different biomarkers. In the last 2 columns we list the hazard ratios (HR) and the p-value when using un-dichotomized, mean-pooled Pos scores for each marker. In the first 4 columns, we list the hazard ratios and p-values that result when using a survival tree to pick a dichotomized pooled staining scores for each marker. For a given biomarker, we used the deviance residuals as outcome in a regression tree that contained all staining scores. Then we picked the primary splitter of the root node and used the corresponding dichotomized covariate in a univariate Cox regression model. Clearly, this overfits the data and all of the resulting p-values are significant at level 0.05. As mentioned above, one can use the formula given in Altman et al. (1994) to correct the p-value for the fact that an optimal cut-off was chosen by the survival tree. Note that the corrected p-values are far less significant than the uncorrected p-values. Incidentally, the p-values were not corrected for the fact that the survival tree picked an optimal staining score (Pos, Max, or PosMax).

Table 2 presents the result for all 304 patients with complete information in these pathology variables. All other clinical variables are significant except *grade*. Based on these variables, we construct three prognostic models: a clinical model which consists of pathology variables only, a biomarker model which contains only biomarkers, and a biomarker/clinical model which is presented in table 2.

The C-index and R^2 values of different Cox regression models are presented in table

3. We used the un-dichotomized, mean-pooled Pos staining score of each marker. To correct for over-fitting due to including multiple covariates in the model, we used the nonparametric bootstrap method implemented in the *validate* function (Harrel, 2001). We chose 300 bootstrap samples. Note that the C-index and the R^2 measure lead to a similar conclusion: the clinical covariates contain more predictive information than the biomarker covariates. However, the combined model that includes both biomarker and clincial information leads to the highest C-index and R^2 value. To test whether the C-index of the combined model is significantly higher than that of the clinical model, we used the *rcorrp.cens* function in the Design library of R (Harrel, 2001). Table 4 lists the U-statistics for comparing different Cox models. Note that the the clinical/biomarker model is significantly better than the clinical model (p-value = 0.00019). Thus, we conclude that the biomarkers provide additional predictive information for survival beyond the clinical predictors. A nomogram based on the biomarker/clinical model is presented in figure 3.

We used the R function rpart to construct regression trees. When dealing with multiple biomarkers, one may use a survival tree to detect possible interactions between them, see figure 4. As pointed out above, the tree method may be too greedy to detect significant interactions, which is why we also analyzed the data with bump hunting.

The results of bump hunting are presented in table 5. We find 2 rules for characterizing high risk patients. Specifically, the data set was randomly divided into a training (2/3 of the data) and a test set (1/3 of the data). Bump hunting was applied to the training data set and the resulting rules evaluated on the test data set. Table 5 lists the mean values of the deviance residuals and the box supports, i.e., the proportion of cases that satisfy the rule. Boxes (rules) B_1 and B_2 cover 35.5% of the initial data set, and only 5% of patients are covered by both sets of rules. Figure 5 shows the Kaplan Meier curves for patient groups defined by different rules. Note that the logrank p-value ($p = 7.0 \times 10^{-8}$) is very small. Since overfitting took place, it should only be considered as descriptive measure of curve separation. The rules suggest how to form interaction terms of biomarker expressions. The significance of the corresponding interaction terms should be tested on future data sets.

6 DISCUSSION

We present statistical tools for addressing important data analysis challenges of TMA data. A fundamental question is whether biomarkers can replace or complement standard clinical and pathology covariates. We discuss statistical methods for evaluating whether biomarkers contain more predictive information than standard clinical or pathology covariates. This is fairly straightforward when the biomarker expression covariate is defined without peaking at the clinical data.

But for novel biomarkers it is often unclear how to score them and how to pool several spot measurements. Biomarker staining scores are challenging since they can be highly skewed, semi-continous, and highly correlated. Biologists typically prefer dichotomized biomarker expression values since they lend themselves to easy biological interpretation: dichotomized biomarker expressions can be interpreted as protein over-expression or protein function loss. This suggests the use of rule induction methods that dichotomize covariates. Survival tree predictors are an obvious choice. We also adapted bump hunting to survival outcomes through the use of deviance residuals. In simulation studies we find that this approach works well. A simulation report can be downloaded from `www.genetics.ucla.edu/labs/horvath/technicalreports`.

In our kidney cancer TMA data analysis, bump hunting works well to detect higher level biomarker interaction effects. However, these should be validated on independent, external test sets. Survival trees and bump hunting are model-free methods that handle non-linear relationships well and lead to easily interpretable results. Rules that characterize high risk patients can be encoded in binary covariates. These should not be tested for significance in Cox regression models that use the same survival data. But for the special situation of finding optimal cut-off values, we review some strategies for correcting the resulting p-values.

To validate biomarker rules, we recommend to (repeatedly) split the data into training and test data. The rules should be constructed on the training data and validated on the test data. Alternatively one may use a data resampling scheme or cross-validation to arrive at unbiased estimates of rule performance.

The increasing use of tissue microarrays for validating tumor markers provides motivation for the development of appropriate statistical methods. Tissue microarray data are very different from DNA gene expression microarray data and require different analysis methods.

ACKNOWLEDGEMENT

We would like to thank Tao Shi and Zixing Fang for interesting discussions. The authors thank the anonymous reviewers for valuable comments on this article.

REFERENCES

Aaronson, K.D., Schwartz, J.S., Chen, T.M., Wong, K.L., Goin, J.E., and Mancini, D.M. Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation. *Circulation*, **1997**, 95(12), 2660–2667.

Ahn, H. and Loh, W.-Y. Tree-structured proportional hazards regression modeling. *Biometrics*, **1994**, 50, 471–485.

Altman, D.G., Lausen, B., Sauerbrei, W. and Schumacher, M. Dangers of using ‘optimal’ cutpoints in the evaluation of prognostic factors. *Journal of National Cancer Institute*, **1994**, 86, 829–835.

Bettencourt M.C., Bauer J.J., Sesterhenn, I.A., Mostofi F.K., McLeod D.G., Moul J.W. Ki-67 expression is a prognostic marker of prostate cancer recurrence after radical prostatectomy. *Journal of Urology*, **1996**, 156, 1064–1068.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. Classification and Regression Trees, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, **1984**.

Bui, M.H., Seligson, D., Han, K.R., Pantuck, A.J., Dorey, F.J., Huang, Y., Horvath, S., Leibovich, B.C., Chopra, S., Liao, S.Y., Stanbridge, E., Lerman, M.I., Palotie, A., Figlin, R.A., Belldegrun, A.S. Carbonic anhydrase IX is an independent predictor of survival in advanced renal clear cell carcinoma: implications for prognosis and therapy. *Clinical Cancer Research*, **2003**, 9, 802–811.

Clark, T.G., Bradburn, M.J., Love, S.B., and Altman, D.G. Survival Analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, **2003**, 89, 781–786.

Ciampi, A., Thiffault, J., Nakache, J.P., and Asselain, B. Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis*, **1986**, 4, 185–204.

Cox, D.R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, **1972**, 34, 187–220.

Efron, B. Estimating the error rate of a prediction rule: Improvements on cross-validation. *Journal of the American Statistical Association*, **1983**,78, 316–331.

Friedman, J. and Fisher, N. Bump hunting in high dimensional data. *Statistical Computing*, **1999**, 9, 123-143.

Hachamovitch, R., Hayes, S.W., Friedman, J.D., Cohen, I., and Berman, D.S. Comparison of the short-term survival benefit associated with revascularization compared with medical therapy in patients with no prior coronary artery disease undergoing stress myocardial perfusion single photon emission computed tomography. *Circulation*, **2003**, 107(23), 2900–2907.

Harrel, F.E. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. *Springer series in statistics*, **2001**, ISBN 0-387-95232-2.

Harrel, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., and Rosati, R.A. Evaluating the yield of medical tests. *Journal of American Medical Association*, **1982**, 247, 2543–2546.

Harrel, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, **1984** , 3(2), 143–152.

Heinzl, H. Dangers of using ‘optimal’ cutpoints in the evaluation of cyclical prognostic factors. *New Approaches in Applied Statistics*, **2000**, 16, 135–143.

Ihaka R. and Gentleman, R. R: a language for data analysis and graphics. *J. Comput. Graphical Statistics*, **1996**, 5, 299–314.

Kaplan, E.L. and Meier, P. Nonparametric estimation from incomplete observations. *J.of the American Statistical Association*, **1958**, 53, 457–48.

Kononen, J., Bebendorf, L., Kallioniemi, A., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M., Sauter, G., and Kallioniemi, O.P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine* **1998**,4, 844–847.

LeBlanc, M. and Crowley, J. Relative risk regression trees for censored survival data. *Biometrics*, **1992**, 48(2), 411–425.

Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, **1991**, 78, 691–692.

Pantuck, A. J., Zisman, A., Belldegrun, A. S. The Changing Natural History of Renal Cell Carcinoma. *Journal of Urology*, **2001**, 166, 1611–2001.

Schoenfeld, D.A. Partial residuals for the proportional hazards regression model. *Biometrika*, **1982**, 69, 239–241.

Schumacher, M., Holländer, N. and Sauerbrei, W. Resampling and cross-validation techniques: A tool to reduce bias caused by model building. *Statistics in Medicine*, **1997**, 16, 2813–2827.

Segal, M. Regression trees for censored data. *Biometrics*, **1988**, 44, 35–48.

Somers, R.H. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, **1962**, 27, 799–811.

Therneau, T., Grambsch, P. and Fleming, T. Martingale based residual for survival models. *Biometrika*, **1990**, 77, 147–160.

Zhang, H. and Singer, B. Recursive Partitioning in the Health Sciences. Springer-Verlag New York, **1999**.

Table 1: Univariate Cox regression models involving biomarkers.

Bio-	Staining score and pooling picked by survival tree				Fixed score(Pos) & mean pool	
marker	score(pooling)	cut-off	HR[95%CI]	p-val. (corrected p)	HR[95%CI]	p-value
CA9	PosMax(min)	<25	2.09[1.49,2.93]	2.0E-05(8.4E-04)	0.994[0.99,0.999]	1.70E-02
P53	Max(mn)	≥ 1.4	2.8[1.85,4.23]	9.7E-07(5.3E-05)	1.02[1.01,1.03]	8.80E-05
Vimentin	Max(mn)	< 1.8	1.84[1.31,2.57]	0.00038(0.011)	1.01[1.00,1.01]	0.067
PTEN	Pos(min)	<60	1.67[1.19,2.34]	0.003(0.063)	0.996 [0.99,1.00]	0.2
Gelsolin	Max(max)	>0	2.03[1.38,2.98]	0.00031(0.009)	1.00[0.997,1.01]	0.46
Epcam	Pos(md)	<5	1.68[1.17,2.41]	0.0046(0.089)	0.992[0.984,1.00]	0.042
CA12	Pos(min)	<80	2.1[1.38,3.2]	0.00055 (0.015)	0.99[0.984,0.996]	0.0018
Ki67	Pos(md)	<15	2.78[1.95,3.97]	1.60E-08(1.15E-06)	1.04[1.03,1.06]	1.10E-07

Table 2: Multivariable Cox regression with undichotomized staining scores and pathology covariates.

Marker	p-value	HR [95%CI]
CA9(Pos.mn)	5.2E-05	0.989[0.984,0.995]
p53(Pos.mn)	1.3E-02	1.011[1.002,1.020]
pTEN(Pos.mn)	2.1E-02	0.993[0.987,0.999]
Vimentin(Pos.mn)	1.7E-02	1.008[1.001,1.014]
met	3.8E-10	4.654[2.877,7.531]
T-stage	2.3E-04	1.560[1.231,1.976]
ECOG	1.8E-04	1.809[1.327,2.467]
grade	1.6E-01	1.196[0.933,1.534]

Table 3: Validated model fitting indices for different Cox models. The bootstrap was used to estimate the optimism of each index.

	Bias-corrected C-index	Bias-corrected R^2
clinical	0.795	0.3764
biomarker	0.615	0.053
biomarker/clinical	0.804	0.415

Table 4: Results of using the *rcorrp.cens* function to compare different Cox models.

Model 1 versus model 2	Model 1 [a]	Model 2 [b]	p-value [c]
clinical versus biomarker	38	5	<0.001
biomarker/clinical versus biomarker	40	3	<0.001
biomarker/clinical versus clinical	8	4	0.00019

[a] percent of patient pairs where model 1 correctly predicts outcome while model 2 does not

[b] percent of patient pairs where model 2 correctly predicts outcome while model 1 does not

[c] U-Statistic p-value

Table 5: Rules produced by bump hunting.

box	data set	box support	box (global) mean	rule
B_1	training	0.2822	1.62(1.29)	Gelsolin(Max.mn)> 0 & P53(Pos.max)> 0.31 & Vimentin(Max.mn)> 1.42
	test	0.2255	1.97(1.36)	
	all	0.2632	1.73(1.31)	
B_2	training-B_1	0.1089	2.05 (1.15)	CA9(PosMax.min)< 27.5 & Ki67(Pos.mn)> 1.77 & Gelsolin(Pos.min)< 15
	test-B_1	0.0882	2.12 (1.17)	
	all-B_1	0.1513	2.07 (1.16)	

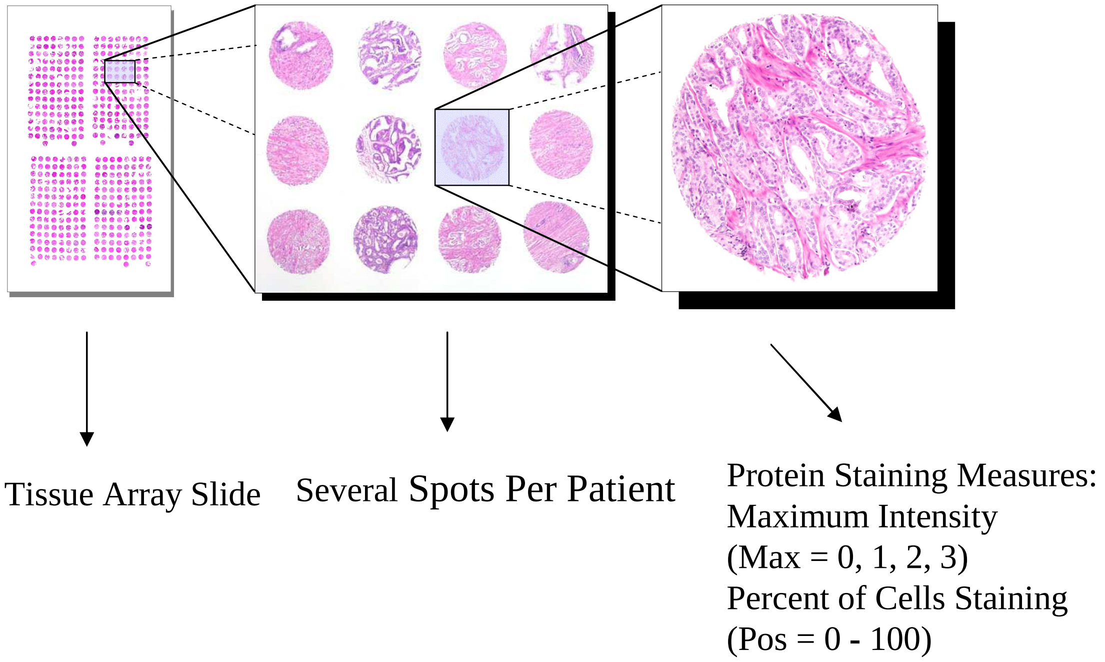
Tissue Array Slide
Several Spots Per Patient
Protein Staining Measures:
Maximum Intensity
(Max = 0, 1, 2, 3)
Percent of Cells Staining
(Pos = 0 - 100)

Figure 1: *A tissue array slide. Each tumor is represented by multiple spots. Several protein staining scores are measured for each spot, e.g. the maximum intensity (Max*$\in \{0, 1, 2, 3\}$*), the percentage of cells staining (Pos* $\in \{0, \ldots, 100\}$*), etc.*

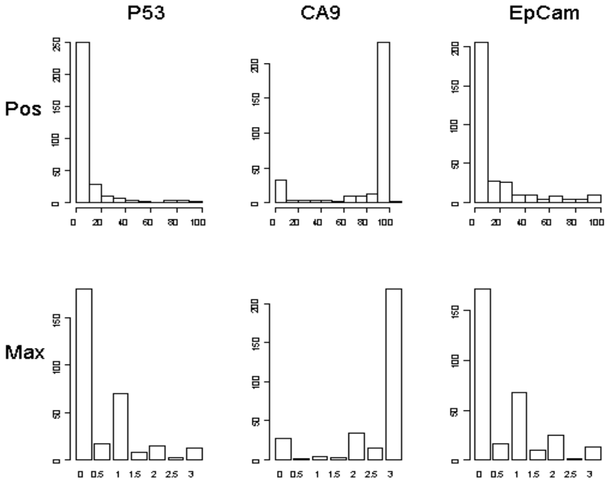

Figure 2: *Histograms of four pooled biomarker staininned scores. We used mean pooling for the Pos scores and median pooling for the Max scores. Note that the Pos scores are highly skewed percentages. The staining intensity (Max) is an ordinal variables with values between* 0 *and* 3.

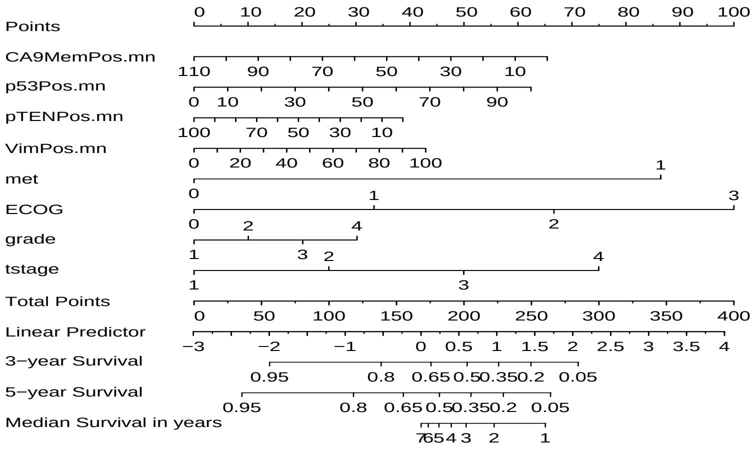

Figure 3: *A prognostic nomogram for the biomarker/clinical model. For each predictor, read the points assigned on the* $0-100$ *scale and add these points. Read the results on the 'Total Points' scale and then read the corresponding predictions below it.*

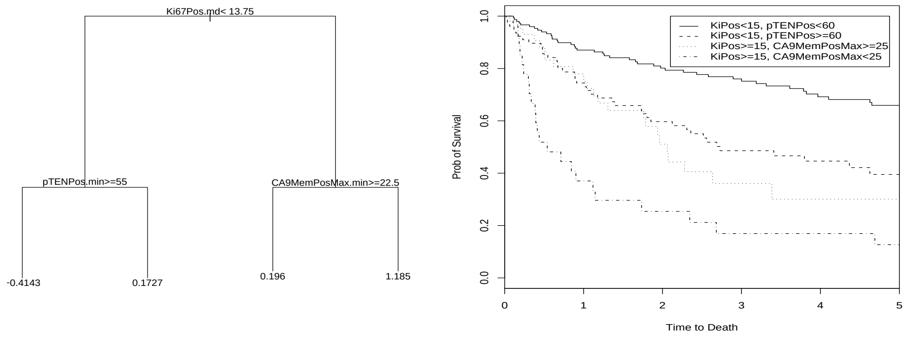

Figure 4: *A survival tree (left panel) which involves multiple biomarkers. It was constructed by fitting a regression tree to the deviance residuals of an intercept only Cox regression model. The higher the number at the terminal note, the worse is the prognosis. Kaplan-Meier curves (right panel) corresponding to the 4 terminal nodes.*

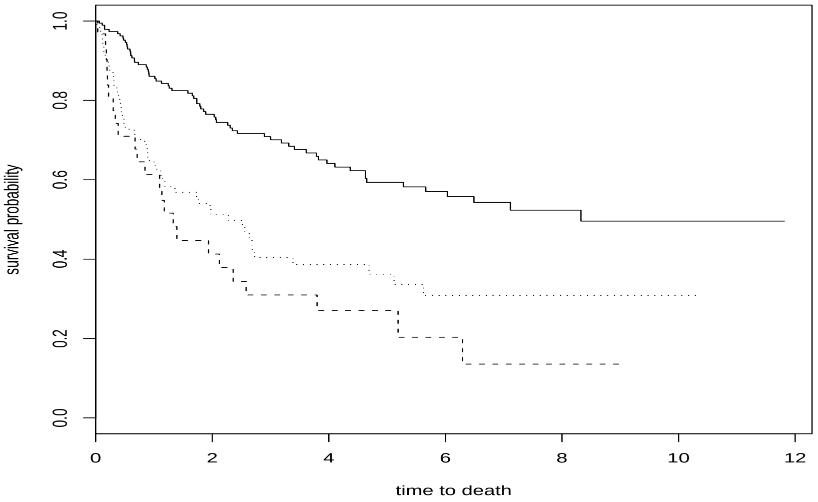

Figure 5: *Kaplan-Meier (KM) curves corresponding to rules resulting from bump hunting. The KM curves of patients who satisfy rule 1, rule 2 but not rule 1, neither rule, are dashed, dotted, and solid, respectively. The logrank test p-value of* $7.08E-8$ *should not be used for inference due to over-fitting: the rules were constructed using the survival information.*