BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 1 2009, pages 54–60
doi:10.1093/bioinformatics/btn354

Genetics and population analysis

The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis

James A. Koziol[1], Anne C. Feng[1], Zhenyu Jia[2], Yipeng Wang[2,3], Seven Goodison[4], Michael McClelland[3] and Dan Mercola[2,*]

[1]The Scripps Research Institute, La Jolla, [2]Translational Cancer Biology, Department of Pathology and Laboratory Medicine, University of California, Irvine, [3]The Sidney Kimmel Cancer Center, San Diego, CA and [4]Department of Surgery, University of Florida, Shands Health Science Center, Jacksonville, FL, USA

Received on April 3, 2008; revised on July 9, 2008; accepted on July 10, 2008

Advance Access publication July 15, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Classification and regression trees have long been used for cancer diagnosis and prognosis. Nevertheless, instability and variable selection bias, as well as overfitting, are well-known problems of tree-based methods. In this article, we investigate whether ensemble tree classifiers can ameliorate these difficulties, using data from two recent studies of radical prostatectomy in prostate cancer.

Results: Using time to progression following prostatectomy as the relevant clinical endpoint, we found that ensemble tree classifiers robustly and reproducibly identified three subgroups of patients in the two clinical datasets: non-progressors, early progressors and late progressors. Moreover, the consensus classifications were independent predictors of time to progression compared to known clinical prognostic factors.

Contact: dmercola@uci.edu

1 INTRODUCTION AND SUMMARY

Classification and regression trees (CART; Breiman *et al*., 1984) have long been used for cancer diagnosis and prognosis (Dillman and Koziol, 1983; Koziol *et al*., 2003). Extensions of the CART methodology to survival analyses (Gordon and Olshen, 1985; LeBlanc and Crowley, 1992; Segal, 1988) are readily available for prediction of survival probabilities. Nevertheless, instability and variable selection bias, as well as overfitting, are well-known problems of tree-based methods. In this article, we investigate whether ensemble tree classifiers (Bühlmann, 2004) can ameliorate these difficulties. In related contexts, ensemble techniques tend to reduce error variances and increase the robustness of findings (Bhanot *et al*., 2006); we might therefore hope that combining the results of several individual trees will yield results more reliable and potentially less biased than any particular tree.

For clarity, we consider a concrete problem, prediction of times to progression with data from two recent studies of radical prostatectomy in prostate cancer (Stephenson *et al*., 2005; Yu *et al*., 2004). Independently of these studies, we have five mutually exclusive gene lists of cardinality 23–100, each putatively associated with prostate cancer diagnosis or prognosis. For each gene list, we extracted gene expression values from Affymetrix microarray analyses performed on biopsy samples from all patients in the two clinical studies. We then determined individual regression trees modeling time to progression from each gene list after filtering, and combined the individual trees into consensus classifiers. In multivariable Cox proportional hazards regression models, the consensus classifiers were significant predictors of time to progression independently of other prognostic factors. Moreover, trichotomous classification of the study samples into non-progressors, early, and late progressors was a robust finding in our unsupervised learning context.

2 METHODS

2.1 Patient cohorts

We have two independent clinical datasets of prostate cancer patients who had undergone radical prostatectomy: Set 1 (Stephenson *et al*., 2005) comprises 79 patients, and Set 2 (Yu *et al*., 2004), 49 patients. Follow-up information was available from all patients post-surgery, with the primary outcome of interest being time to biochemical progression, defined as a recurrence of a detectable PSA value after prostatectomy.

2.2 Gene lists

We assembled five exclusive lists of genes, here designated SKCC1 through SKCC5, each consisting of up to 100 entries, and putatively associated with prostate cancer diagnosis or prognosis. SKCC1 consisted of 23 prostate cancer-related genes extracted from the literature by one of the authors (D.M.); SKCC2 consisted of 100 tumor-specific relapse-related genes, and SKCC3, 100 stroma-specific relapse-related genes, both from a follow-up study to Stuart *et al*. (2004); SKCC4 consisted of 33 relapse-related genes from Glinsky *et al*. (2005); and SKCC5 consisted of 88 genes related to Gleason score from True *et al*. (2006). Our working hypothesis is that the genes or perhaps a subset of the genes in each list are 'predictive' of progression in Set 1 and Set 2. Note that we are not screening through thousands of genes in an exploratory mode; hence our statistical approach can be more focused on assessing the underlying hypothesis.

2.3 Statistical methods

2.3.1 Microarray data Set 1 contained Affymetrix U133A array data from 79 patients, and Set 2 contained Affymetrix U95Av2/U95B/U95C array data from 48 patients. Affymetrix array .CEL files were processed using the *R* statistical language; in particular, the RMA algorithm from the 'affy'

*To whom correspondence should be addressed.

package (http://www.bioconductor.org) was used for data normalization. Microarray data from different platforms was normalized separately. These datasets are available from one of the authors (D.M.).

For each gene list, and each clinical dataset (Set 1 and Set 2), we extracted the relevant expression values from the normalized microarray dataset corresponding to Set 1 or Set 2, and then filtered the gene list to identify those genes that appeared to be differentially expressed. For this purpose, we utilized EMMIX, as described previously (McLachlan *et al*., 2002). EMMIX attempts to fit mixtures of t distributions to the gene expression patterns across each cohort, with the implication that a mixture distribution is indicative of potential discriminatory power for that gene. Specifically, we utilized EMMIX for 10 different datasets [5 gene lists $\times$ 2 microarray datasets]. In each EMMIX run, the software automatically assessed the relevance of each of the N genes by fitting one- and two-component t mixture models to the expression data over the 79 tissue samples (Set 1) or 49 tissue samples (Set 2) for each gene considered individually. The decision of whether a gene is 'interesting' or not is made on the basis of whether the likelihood ratio statistic for testing one versus two components in the mixture model exceeds a prespecified threshold. [Following McLachlan *et al*. (2002), we chose a threshold of 8 for minus twice the log likelihood ratio. We must emphasize, however, that there is no attempt to impute statistical significance to this choice of threshold.] We thereby achieved reductions in gene lists ranging from 66.7% to 82.1% [median 73.9%] for Set 1, and ranging from 67.9% to 91.6% [median 71.4%] for Set 2. We subsequently used the gene expression levels from these remaining 'interesting' genes as input for constructing survival trees, with time to progression as the relevant outcome variable.

2.3.2 Time-to-progression data Kaplan–Meier curves were used to summarize times to progression within patient subgroups, and logrank statistics were used to compare times to progression between subgroups. Two-sided P-values are reported for the logrank statistics; in this regard, P-values for the logrank statistics applied to the subgroups identified by the trees were estimated from 10 000 random permutations of the underlying samples rather than from putative asymptotic distribution theory. Separation between time-to-progression curves was summarized with Efron's version of Harrell's concordance index C (Efron, 1967; Harrell *et al*., 1982; Koziol and Jia, 2007). [Recall, from Efron, that, if F and G denote independent survival distributions, then the concordance C between F and G is Prob$[X > Y]$, where $X \sim F$ and independently $Y \sim G$. In the setting of randomly censored survival data, Harrell's concordance index does not consistently estimate the Mann–Whitney parameter Prob$[X > Y]$, whereas Efron's does. See Koziol and Jia (2007) for further details].

2.3.3 Decision trees We used the *rpart* routines of the R language (www.r-project.org) to develop survival trees based on the gene expression data. We based splits on deviance, pruned the trees via complexity parameters to a maximum of five terminal nodes, and utilized 10-fold cross-validation to avoid overfitting the training data. Nevertheless, as noted by others (Tsai *et al*., 2007), the terminal subgroups of patients may still have similar profiles of times to progression. We therefore augmented the usual tree pruning methodology by comparison of times to progression via logrank statistics, and grouping the closest (that is, statistically 'non-significant', using α-level 0.05 cutoff) pairs of subgroups together relative to this measure of similarity. Tsai *et al*. (2007) have shown that this approach tends to perform more efficiently than standard tree pruning methods for producing homogeneous groupings of patients. We invariably achieved tripartite classification schemes with each of the trees: if trees had three terminal nodes, each of the survival curves was putatively significantly different from the others; if trees had four or five terminal nodes, homogeneity was achieved by pooling the 'non-extreme' cohorts on the KM curves.

2.3.4 Ensemble classifiers A recognized limitation of decision trees is that they can be somewhat sensitive to input characteristics, leading to overfitting. Ensemble methods are learning algorithms that construct a set of base classifiers and then classify new data points by taking a vote of their predictions. Notable examples involving tree-based classification methods include bagging (Breiman, 1996), boosting (Freund and Schapire, 1997) or arcing (Breiman, 1998): typically with bagging, boosting and arcing, the combined classifier enjoys improved operating characteristics compared to each of the individual base decision trees. We here adopted a straightforward approach, simple ensemble voting, whereby classification is implemented by unweighted voting among the individual component trees. Briefly, each of the five individual trees (one tree per gene list) classifies a patient into one of three ordered subgroups that manifest low, intermediate or high risk of progression relative to one another; from these five trees, we determined the consensus classification by unweighted averaging of the classifications across the individual trees.

2.3.5 Prognostic importance We utilized multivariable Cox regression analyses to assess the prognostic importance of the clinical factors Gleason score, TNM classification and preoperative PSA levels, as well as the consensus classification schemes, on times to relapse with Sets 1 and 2.

2.3.6 Diversity of the individual trees We computed weighted κ-statistics (Dietterich, 2000; Fleiss, 1981); as measures of diversity among the various trees. The κ's were calculated from the coincidence matrices of each pair of trees, these being derived from the tripartite classification schemes from each tree, using square weighting on the indices of the categories.

3 RESULTS

3.1 Patient characteristics

Clinical characteristics of the two cohorts are summarized in Figure 1a–c. The two sets do not differ significantly relative to Gleason scores (Fisher test, $P = 0.69$) or preoperative PSA values (Fisher test, $P = 0.26$). On the other hand, patients in Set 2 were classified with more severe TNM staging than those in Set 1 (Fisher test, $P < 10^{-10}$). In Figure 1d, we present the Kaplan–Meier time-to-progression curves for the two cohorts: clearly, the two cohorts are somewhat disparate, with early progression a much more common event with Set 2 compared to Set 1.

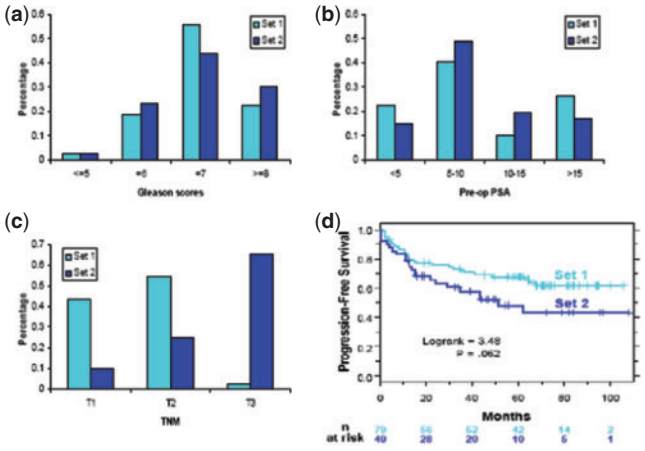

Fig. 1. Clinical characteristics of two cohorts of prostate cancer patients, of sizes 79 (Set 1) and 49 (Set 2), respectively. (**a**) Proportionate representation of Gleason scores. (**b**) Proportionate representation of Preoperative PSA levels. (**c**) Proportionate representation of TNM values. (**d**) Progression-free survival curves. Numbers of subjects at risk in the two sets at various time points are indicated beneath the x-axis of the Kaplan–Meier plot.

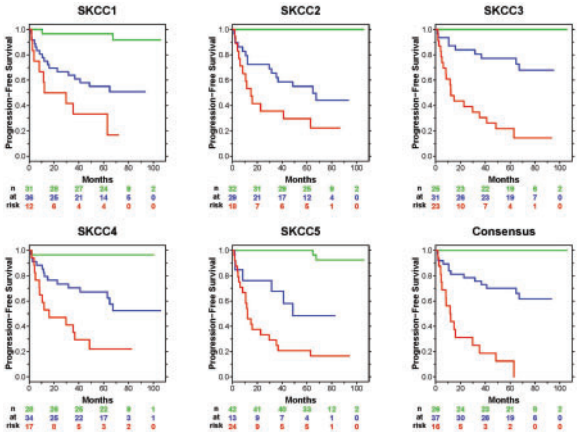

Fig. 2. Progression-free survival curves from the tree-based classifiers based on gene lists SKCC1 through SKCC5, and the consensus classifier derived from the ensemble, for Set 1. Numbers of subjects at risk in the three subgroups at the various time points are indicated beneath the *x*-axis of each graph.

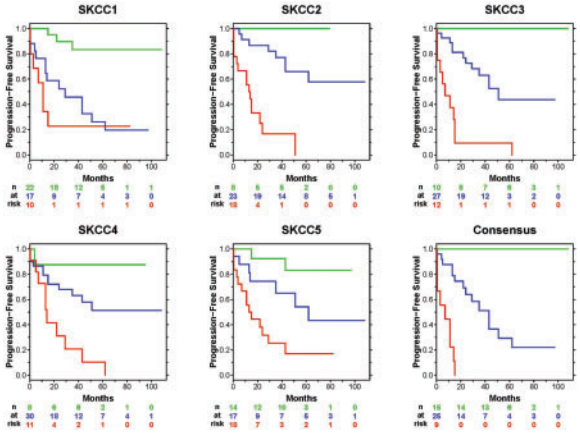

Fig. 3. Progression-free survival curves from the tree-based classifiers based on gene lists SKCC1 through SKCC5, and the consensus classifier derived from the ensemble, for Set 2. Numbers of subjects at risk in the three subgroups at the various time points are indicated beneath the *x*-axis of each graph.

3.2 Time-to-progression trees

We utilized McLachlan's EMMIX procedure (McLachlan *et al.*, 2002) to determine which of the genes in each gene list were differentially expressed in the Set 1 and Set 2 cohorts. We then used these gene expression levels as input for determining survival trees (with time-to-progression as the relevant outcome variable). Following pruning and grouping, the time-to-progression curves derived from each tree are shown in Figures 2 and 3. Note that the trees generally determine three ordered 'subsets' of patients, in terms of times to progression, regardless of input set. Since trees can be somewhat sensitive to input characteristics, we next attempted to provide a more stable classifier, using the notion of ensemble voting as described above. The survival curves derived from the consensus classifiers are also given in the figures.

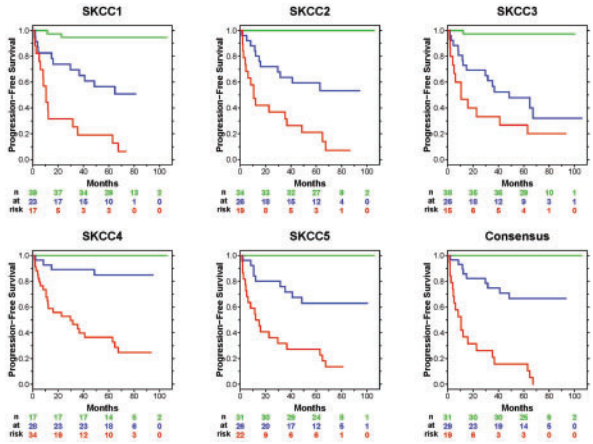

Fig. 4. Progression-free survival curves for Set 1. Here, the tree-based classifiers were derived from the genes highlighted in Set 2 rather than from Set 1 (as in Fig. 1); the consensus classifier was derived from the ensemble of these five classifiers. Numbers of subjects at risk in the three subgroups at the various time points are indicated beneath the *x*-axis of each graph.

In this initial analysis, we developed trees for the two datasets that reveal different levels (risks) of progression. We undertook another investigation of stability of consensus classifiers, via cross-training. From each gene list, we took the genes that passed the EMMIX screening from Set 1 (resp., Set 2), and used these to develop trees with the Set 2 (resp., Set 1) cohort. We then again established consensus classifiers, utilizing the procedure described previously. The individual trees are somewhat noisier than those developed from the training sets, but the consensus classifiers are quite similar (Figs 3 and 4).

We remark that the most consistent finding with all of the trees is the clear identification of a subgroup of patients who do not experience progression. Interestingly, with Set 1, 26 of the 79 patients are classified as non-progressors with the original consensus classifier (Fig. 2); the non-progressor cohort expands to 31 of the 79 patients from the cross-training consensus classifier (Fig. 4). In contrast, with Set 2, 15 of 49 patients are classified as non-progressors with the original consensus classifier (Fig. 3); this drops to 10 of 49 with the cross-training consensus classifier (Fig. 5).

3.3 Concordance with the consensus classifiers

We calculated estimates (and associated standard errors) of the concordance *C* between the consensus tree-derived survival curves for the low- , intermediate- and high-risk subgroups. The low-risk subgroups are clearly differentiated from the intermediate- and high-risk subgroups in terms of concordance. The concordance between the intermediate- and high-risk subgroups is not so pronounced, but still never falls below 0.8 (Table 1).

3.4 Comparison of the consensus classifiers and clinical factors

We used multivariable Cox models to examine whether the consensus classifiers were of prognostic importance relative to the clinical risk factors Gleason score, TNM classification and preoperative PSA levels. Results are summarized in Table 2. Results for Set 1 and Set 2 are similar: of the three clinical risk

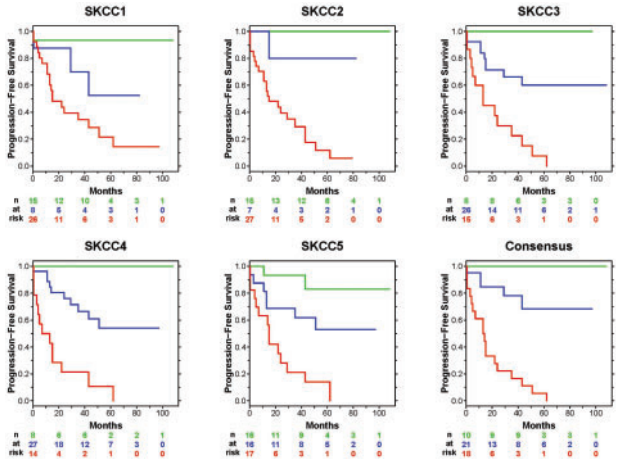

Fig. 5. Progression-free survival curves for Set 2. Here, the tree-based classifiers were derived from the genes highlighted in Set 1 rather than from Set 2 (as in Fig. 2); the consensus classifier was derived from the ensemble of these five classifiers. Numbers of subjects at risk in the three subgroups at the various time points are indicated beneath the *x*-axis of each graph.

Table 1. Concordance indices for the four consensus classifiers

Classifier	L versus H	L versus M	M versus H
Set 1 original	0.94 (0.013)	0.90 (0.014)	0.81 (0.016)
Set 1 validation	0.98 (0.004)	0.89 (0.022)	0.82 (0.014)
Set 1 original	1 (0)	1 (0.049)	0.90 (0.016)
Set 1 validation	1 (0)	1 (0.241)	0.86 (0.025)

For each classifier, L denotes the low-risk subgroup, M the intermediate-risk subgroup, and H, the high-risk subgroup. Tabulated values are Efron's concordance indices calculated from the survival curves presented in Figures 2–5; values in parentheses are bootstrap standard errors of these estimates.

factors Gleason scores, TNM classification and preoperative PSA levels, only Gleason scores are individually significant predictors of relapse. Regardless, the consensus classifiers are stronger predictors of relapse than Gleason scores. In the multivariable setting, the consensus classifiers tend to outperform the combination of Gleason, TNM and preoperative PSA. This can be easily deduced on the basis of a likelihood ratio test for nested models. For example, the prognostic contribution of C_o to a model that already contains G, T and P for Set 1 can be assessed by looking into the χ^2-statistic 31.1 (54.8 to 23.7) with 2 (10 to 8) degrees of freedom, which is highly significant with P-value = 1.76e-07. In addition, the G+C models constitute a significant improvement in fit over G alone, but not over C alone. Nor do the G+T+P+C models provide a significantly improved fit over G+C alone.

3.5 Diversity of the trees

Diversity is an important characteristic in classifier combination (Cunningham and Carney, 2000). Indeed, an ensemble classifier can be more accurate than its component classifiers only if the individual classifiers disagree with one another (Hansen and Salamon, 1990). Various measures of diversity in classifier ensembles have been proposed (Kuncheva and Whitaker, 2003); we selected a well-known measure, the weighted κ-statistic (Dietterich, 2000; Fleiss, 1981);

Table 2. Cox regressions comparing clinical risk factors and the consensus classifiers as predictors of relapse

	Predictors	R^2	LRT	(df)	P-value
Set 1					
	G	0.19	16.2	(3)	1.03e-03
	T	0.06	4.5	(2)	0.11
	P	0.06	5.0	(3)	0.17
	G+T	0.23	20.9	(5)	S.40e-03
	G+P	0.21	18.4	(6)	0.0054
	T+P	0.14	11.6	(5)	0.041
	G+T+P	0.26	23.7	(8)	2.58e-03
	C_o	0.35	34.4	(2)	3.43e-08
	C_v	0.44	45.4	(2)	1.37e-10
	G+C_o	0.40	39.9	(5)	1.58e-07
	G+C_v	0.45	47.8	(5)	3.88e-09
	G+T+P+C_o	0.50	54.8	(10)	3.52e-08
	G+T+P+C_v	0.53	59.4	(10)	4.73e-09
Set 2					
	G	0.21	10.1	(3)	0.018
	T	0.06	2.4	(2)	0.31
	P	0.06	2.7	(3)	0.44
	G+T	0.23	10.5	(5)	0.062
	G+P	0.30	13.1	(6)	0.042
	T+P	0.11	3.8	(5)	0.58
	G+T+P	0.30	12.3	(8)	0.14
	C_o	0.59	44.0	(2)	2.74e-10
	C_v	0.51	34.6	(2)	3.12e-08
	G+C_o	0.61	40.8	(5)	1.05e-07
	G+C_v	0.54	33.2	(5)	3.37e-06
	G+T+P+C_o	0.66	36.7	(10)	6.47e-05
	G+T+P+C_v	0.61	31.8	(10)	4.29e-04

The rows correspond to Cox regressions with the indicated predictors included in the model specification. The predictors are denoted G, Gleason scores, T, TNM classification, P, pre-operative PSA levels, and C, consensus classifier. G, T, and P are categorical variables with 4, 3 and 4 categories, respectively, as given in Table 1. C_o and C_v are also categorical variables: C_o corresponds to the consensus classifiers derived from the gene lists for the same dataset, and C_v, to the consensus classifiers derived from the gene lists from the other dataset. C_o and C_v each have three categories as shown in the corresponding Kaplan–Meier curves. Generalized R^2 statistics, and likelihood ratio test statistics (denoted LRT) relative to the null model with associated degrees of freedom (df) are given, along with the approximate P-values from the χ^2 approximation.

to summarize the diversity among the various individual trees. The κ-statistics were calculated from the coincidence matrixes of the tripartite classification schemes of each pair of trees. These diversity measures are given in Table 3. In general, the trees tend to be congruent, though pairwise agreement is far from perfect. We remark that the level of agreement increases if we dichotomize the classification schemes of each tree into the non-progressors (the 'flat-liners' in Figs 2–5) versus all others. This is not altogether surprising, as the trees consistently tend to accurately identify a subgroup of patients who do not experience relapses.

4 DISCUSSION

Classification and regression trees have long been used in studies of cancer diagnosis and prognosis (Dillman and Koziol, 1983; Koziol *et al.*, 2003). Nevertheless, a recognized limitation of decision trees is that they can be somewhat sensitive to input characteristics,

Table 3. Diversity measures for the individual trees

A. Set 1

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	0.49	0.74	0.5	0.51
[2,]	0.25	1	0.53	0.42	0.48
[3,]	0.41	0.44	1	0.65	0.59
[4,]	0.57	0.51	0.55	1	0.51
[5,]	0.49	0.56	0.54	0.59	1

B. Set 2

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	0.18	0.45	0.27	0.45
[2,]	0.49	1	0.29	0.15	0.39
[3,]	0.44	0.47	1	0.49	0.22
[4,]	0.52	0.61	0.59	1	0.36
[5,]	0.21	0.47	0.32	0.34	1

The bracketed numbers [i,] and [,j] (row and column indices in each 5×5 matrix above) denote the individual trees derived from gene list SKCCi, $i = 1,2,3,4,5$. The tabulated entries in position (i, j) of each 5×5 matrix are the weighted κ statistics for comparing trees i and j, calculated from the coincidence matrix of the tripartite classification schemes determined from trees i and j. The upper diagonal entries in each matrix correspond to the trees derived from the original gene lists for each dataset, and the lower diagonal (shaded) entries, from the gene lists determined from the other dataset, as detailed in Section 3.1. In other words, the upper diagonal entries in A correspond to the trees depicted in Figure 2, and the lower diagonal en-tries, to the trees depicted in Figure 4. Similarly, the upper diagonal entries in B correspond to the trees depicted in Figure 3, and the lower diagonal entries, to the trees depicted in Figure 5.

leading to overfitting. The purpose of our study was to explore whether this limitation could be overcome by means of ensemble classifiers.

The individual trees developed from the various gene lists rather successfully establish different 'signatures' related to good or poor clinical outcome. Our ensemble method is meant to combine these different signatures into a more robust and accurate prognostic tool. This method is fairly transparent, and we could surely improve on our ensemble classifier by constructing more sophisticated base classifiers (e.g. via bagging, boosting or arcing), or by adopting a weighted voting scheme to combine the individual base classifiers. Nevertheless, it is reassuring that our simple and direct approach does provide tangible evidence of accurate prognoses within our patient cohorts.

In general, addition of clinical or histopathological variables to gene expression levels would be expected to improve classification accuracy. It is somewhat surprising, therefore, that with our datasets, the classical prognostic variables based on the TNM classification scheme, and preoperative PSA levels are not strongly related to patients' clinical courses, in contradistinction to Gleason scores. The consensus tree classifiers are the strongest predictors of outcome in the multivariate Cox regression models that include the standard prognostic factors; in particular, the consensus classifiers do not merely reflect Gleason scores, but remain independent predictors of relapse in the multivariable setting. [As Dunkler *et al*. (2007) have noted, most prediction rules using gene expression have not provided a substantially improved prognostic classification compared with conventional prognostic factors.] In this regard, we remark that nomograms for prediction of progression in prostate cancer, based on preoperative PSA, TNM and Gleason, are of widespread use in urology; our findings suggest that consensus classifiers can bring independent information to bear on nomogram construction and prediction.

Dietterich (2000) has indicated that there is a certain tradeoff between diversity and accuracy for ensemble methods. There is a tradeoff: prediction accuracy of the ensemble classifier that might be expected to increase by adding more classifiers; but improvement might be slight if the classifiers are positively correlated. That is, the ensemble error rate is most reduced in ensembles whose members make individual errors in a less correlated manner. (Hansen and Salamon, 1990; Kuncheva and Whitaker, 2003). In our setting, the individual tress were derived from non-overlapping gene lists, hence one might expect correlations not to be intrinsically high. This diversity was borne out by examination of the weighted κ-statistics (Table 3): few of the pairwise κ's are even so large as 0.6, a level indicative of moderate agreement.

Multiple classifiers can capture various aspects of the underlying biological phenomena; and, combining classifiers can improve operating characteristics. Nevertheless, let us recognize a cogent criticism of ensemble classifiers: namely, the ensemble provides little insight into how it makes its decisions. A single decision tree is easily interpreted; less so with decision trees amalgamated via voting, and even less so with weighted voting.

Let us turn to clinical aspects of our findings. We noted in the results (and a reviewer has presciently observed) that early biochemical progression subsequent to radical prostatectomy was a common occurrence, particularly with Set 2. We deliberately chose not to exclude these patients from our analyses. Previous survival studies of patients following radical prostatectomy have repeatedly demonstrated that the recurrence population of cases experience the most rapid decline of the percent remaining disease free at early times following surgery; rate of recurrence continuously slows with increasing time (negative second derivative for all T). That is, the phenomenon of relapse increases with decreasing times and is maximal immediately after surgery. Hence, patients with positive PSA values in the pos-operative setting are to be expected for the relapse population. Further, given our underlying hypothesis that genotype dictates phenotype, we are equally interested in the genes related to the phenotype of metastasis or spread already at the time of surgery. In other words, the medical challenge is to identify patients that are unlikely to be cured by surgery, and those with spread or metastasis at the time of surgery are among this group.

A reviewer has commented that the selection of our five gene lists 'looks suspicious'. Four of the five gene lists (SKCC1 through SKCC4) were generated 'in-house' by us and our colleagues at the Sidney Kimmel Cancer Center, and were therefore 'natural' candidates for inclusion. The fifth dataset, SKCC5, derived from True *et al*. (2006), was chosen for inclusion because of the exemplary nature of the underlying research: the investigators had used a case set of over 800 cases (by far the largest appearing in the literature to date), and they validated against an independent dataset. Indeed, we are making no claims of prognostic significance of particular gene signatures: rather, we are arguing that consideration of independent predictions (the individual trees) jointly improves the prognosis prediction for individual patients. This notion is not at all unique to us, but is congruent with much previous research. Fan *et al*. (2006), for example, found that multiple gene expression profiles obtained from different laboratories have little overlap in terms of gene identity, but they have high rates of concordance in their outcome predictions for the individual samples. Our five gene sets are mutually exclusive, and have roughly equivalent power to distinguish poor outcomes from good outcomes. A consensus scheme will increase both the accuracy and the precision of classification relative to the individual determinations. We revisit this issue in the Appendix, where we address this reviewer's criticisms in a greater detail.

With regard to the motivating question of prognosis following radical prostatectomy, we have a somewhat ambivalent message: it is possible to construct predictive models using different subsets of genes with putatively different biological mechanisms. One might well argue that redundancy in the information gleaned from the different gene lists seems to be the norm rather than the exception, a cautionary note for researchers interested in establishing classifiers. We emphasize that our goal in this study was not to identify a unique set of genes from which a definitive prediction rule would be based: indeed, the concept of uniqueness in this context is fraught with potential pitfalls (Ein-Dor *et al*., 2004), and lists of genes are highly unstable (Michiels *et al*., 2005). From a statistical perspective, ensemble voting in this context appears to produce a robust and meaningful classifier, and provides a straightforward way of utilizing information from disparate sources. In this regard, we agree strongly with Simon *et al*. (2003, 2004), who argued that the development of a multigene expression profile-based predictor of outcome is a prediction problem, not an inference problem. That is, the objective is accurate prediction, not to identify which genes are associated with outcome, or to ensure that all the genes included in the predictor function are necessary. In general, many genes are correlated, and the genes selected in the model may not be stable under replication or resampling. Hence, although the ensemble methods here build on a set of diverse classifiers, we make no claim of uniqueness of the genes utilized in the individual trees.

On the other hand, we do impute some importance to the fact that the individual trees as well as the ensemble classifiers consistently tend to identify a subgroup of patients that remain progression-free. This raises obvious questions regarding biological and clinical significance. Following Massague (2007), we might speculate that the gene signatures from the individual trees may reflect a common set of phenotypic traits related to non-progression, that is, 'they may be regarded as different pictures of the same beast'. In this spirit, Fan *et al*. (2006) found significant agreement in the outcome classifications of five gene expression-based predictors for breast cancer, and concluded that the different gene sets were probably tracking a common set of biologic phenotypes. Indeed, we might argue that the prognostic worth of the ensemble classifier is the identification of a subset of non-progressors, for whom adjuvant chemotherapy subsequent to prostatectomy might well be unnecessary. There is a continuum of progressors, and the partition of this subset into 'early' and 'late' progressors may constitute a useful but crude summary distinction.

ACKNOWLEDGEMENTS

We thank the reviewers for their insightful comments and suggestions.

Funding: National Institutes of Health (U01CA114810; P01CA1 04898).

Conflict of Interest: none declared.

REFERENCES

Bhanot,G. *et al*. (2006) A robust meta-classification strategy for cancer detection from MS data. *Proteomics*, **6**, 592–604.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Breiman,L. (1998) Arcing classifiers. *Ann. Statist.*, **26**, 801–849.

Breiman,L. *et al*. (1984) *Classification and Regression Trees*. Wadsworth, California.

Bühlmann,P. (2004) Bagging, boosting and ensemble methods. In Gentle JE *et al*. (eds) *Handbook of Computational Statistics*. Springer-Verlag, Berlin. pp. 877–907.

Dietterich,T. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. *Mach. Learn.*, **40**, 139–157.

Dillman,R.O. and Koziol,J.A. (1983) Statistical approach to immunosuppression classification using lymphocyte surface markers and functional assays. *Cancer Res.*, **43**, 417–421.

Dunkler,D. *et al*. (2007) Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *Eur. J. Cancer*, **43**, 745–751.

Efron,B. (1967) The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 4. pp. 831–853.

Ein-Dor,L. *et al*. (2005) Outcome signature genes in breast cancer: is there a unique set. *Bioinformatics*, **21**, 171–178.

Fan,C. *et al*. (2006) Concordance among gene-expression-based predictors for breast cancer. *New Engl J. Med.*, **355**, 560–569.

Fleiss,J. (1981) *Statistical Methods for Rates and Proportions*. John Wiley, New York.

Freund,Y. and Schapire,R. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, **55**, 119–139.

Glinsky,G.V. *et al*. (2005) Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J. Clin. Invest.*, **115**, 1503–1521.

Gordon,L. and Olshen,R. (1985) Tree-structured survival analysis. *Cancer Treat. Rep.*, **69**, 1065–1069.

Hansen,L. and Salamon,P. (1990) Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, **12**, 993–1001.

Harrell,F.E. *et al*. (1982) Evaluating the yield of medical tests. *J. Amer. Med. Assoc.*, **247**, 2543–2546.

Koziol,J.A. and Jia,Z. (2007) The concordance index C with randomly censored data (in press).

Koziol,J.A. *et al*. (2003) Recursive partitioning as an approach to selection of immune markers for tumor diagnosis. *Clin. Cancer Res.*, **9**, 5120–5126.

Kuncheva,L.I. and Whitaker,C.J. (2003) Measures of diversity in classifier ensembles and their relationship to ensemble accuracy. *Mach. Learn.*, **51**, 181–207.

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

Massague,J. (2007) Sorting out breast cancer gene signatures. *New Engl. J. Med.*, **356**, 294–297.

McLachlan,G.J. *et al*. (2002) A mixture model-based approach to the clustering of microarray data. *Bioinformatics*, **18**, 413–422.

Michiels,S. *et al*. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Segal,M.R. (1988) Regression trees for censored data. *Biometrics*, **44**, 35–47.

Simon,R. (2004) An agenda for clinical trials: clinical trials in the genomic era. *Clin. Trials*, **1**, 468–470.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst.*, **95**, 1–14.

Stephenson,A.J. *et al*. (2005) Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, **104**, 290–298.

Stuart,R.O. *et al*. (2004) In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 615–620.

True,L. *et al*. (2006) A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proc. Natl Acad. Sci. USA*, **103**, 10991–10996.

Tsai,C.A. *et al*. (2007) An integrated tree-based classification approach to prognostic grouping with application to localized melanoma patients. *J. Biopharm. Stat.*, **17**, 445–460.

Yu,Y.P. *et al*. (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790–2799.

APPENDIX

You can't make a silk purse out of a sow's ear.
Jonathan Swift (1667–1745)

Jonathan Swift's adage pithily summarizes common wisdom, that it is at best difficult to make something excellent from poor material.

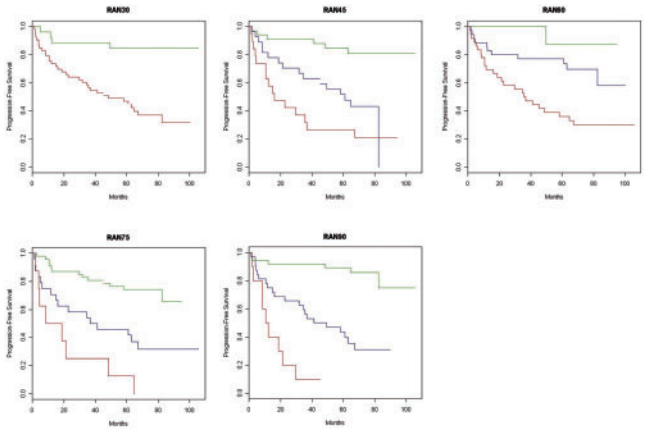

Fig. A1. Progression-free survival curves from the tree-based classifiers based on gene lists RAN30 through RAN90 for Set 1.

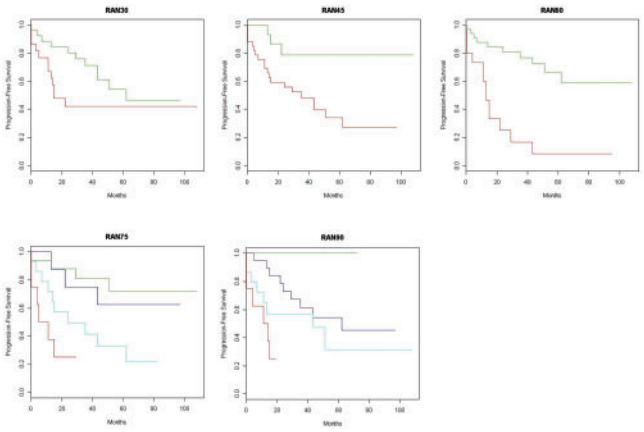

Fig. A2. Progression-free survival curves from the tree-based classifiers based on gene lists RAN30 through RAN90 for Set 2.

We shall investigate in this appendix whether the wisdom of the commons from 300 years ago is maintained in the present. Our purpose here is to respond to one reviewer's injunction: this reviewer has insisted that we repeat our experimental procedure with 'random gene lists'. We remain nonplussed by this demand. As we have noted in the text, our a priori belief with our five gene lists is that each ought to have some discriminatory power in prostate cancer prognosis. The ensemble classifier is then specifically designed to combine relatively weak classifiers into a stronger entity. To begin with random gene lists is to cast us into the realm of exploratory data analysis, in contradistinction to the confirmatory data analytic context under which we were operating in the text. We may 'get lucky' and uncover some genes that are related to prostate cancer prognosis, but this would be at best a hypothesis-generating exercise.

Nevertheless, we will persevere in this endeavor. We begin by drawing five random 'gene lists' of cardinalities 30, 45, 60, 75 and 90, respectively, both from the U133A array data (Set 1, 79 patients), and from the U95 array data (Set 2, 49 patients). (There are some 22 000 probes on the U133A array chip, and 38 000 probes on the U95 array chip.) As before, we filtered the gene lists with EMMIX. Even at this preliminary stage, there is a noteworthy difference with the random gene lists compared with our original gene lists: with the random gene lists, we filtered out 29/30 (96.7%), 40/45 (88.9%), 55/60 (90.2%), 66/75 (88.0%) and 81/90 (90.0%) genes from Set 1, and 29/30 (96.7%), 41/45 (91.1%), 55/60 (90.2%), 70/75 (93.3%) and 86/90 (95.6%) genes from Set 2, far greater proportions than with the original gene lists.

Next, we undertook to partition the patient cohorts into relatively homogeneous subgroups on the basis of progression-free survival, from tree-based classifiers based on the filtered gene lists. Again, our methodology is identical to that described in the text. The resulting progression-free survival curves are shown in Appendix Figures A1 and A2. We believe that the results here are less compelling than our findings with the SKCC gene lists that were presented in the text, in particular, Figures 2 and 3. With both Set 1 and Set 2, there no longer is a clear demarcation into three subgroups with the random gene lists. With Set 1, 30 random genes, we have only two subgroups; and, with Set 2, we have anywhere from two to four subgroups, and separation is compromised. It is not yet altogether transparent how one should construct an ensemble classifier in these circumstances. With both sets, we can no longer distinguish a subgroup of patients who do not experience relapse. This is particularly unfortunate, since there are enormous clinical implications in terms of post-prostatectomy disease management if one could incontrovertibly identify such patients.

In conclusion, we believe Swift's dictum remains relevant today. More fundamentally, however, we remain resolute in the validity of our inferential approach. We derived consensus classifiers that we had a priori hypothesized would have prognostic significance. This hypothesis is testable via Cox regression, hence theoretically is falsifiable. That the consensus classifiers do appear to have prognostic significance independent of the usual clinical variables demonstrates a useful property of ensemble classifiers, but one should not lose sight of the fact that the individual classifiers were based on gene lists that were expected to inform us on the prognostic status of the cohorts. This conceptual paradigm would be vitiated with random gene lists: one might glean some insight into the comparative performance of classifiers with random data, but inference derived therefrom is largely conjectural.

Postscript: the reviewer has also requested 'a comparison to performance of SVM as a gold standard classification procedure'. Whether there exists a gold standard classification procedure in the context of censored survival data is itself a moot point, and would motivate comparison of trees to SVM. Nevertheless, we believe this to be outside the ambit of the present article; we will undertake such comparison, incorporating ensemble classification schemes, in future research.