Biometrics 000, 000–000 DOI: 000
000 0000

Web-based Supplementary Materials for "Greedy Outcome Weighted Tree Learning of Optimal Personalized Treatment Rules" by Ruoqing Zhu, Ying-Qi Zhao, Guanhua Chen, Shuangge Ma and Hongyu Zhao

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A: Tree pruning

We adopt the cost complexity pruning proposed by Breiman et al. (1984), again, injected with subject weights. The cost complexity pruning sets a trade off between the overall misclassification error on the training data with the tree model size through a tuning parameter α. Here, the weighted training data misclassification error for a decision rule $D(\mathbf{X})$ is simply $\frac{1}{\sum w_i}\sum w_i \mathbf{1}_{\{A_i \neq D(\mathbf{X})\}}$. Denote the fully grown tree model $\widehat{D}(\mathbf{X})$ obtained through our algorithm as $\mathcal{T}$, then each internal node in the tree is the starting point for a sub-tree which will end with several leaves (terminal nodes). Denote these sub-trees as $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_M$. Then, a sub-tree $\mathcal{T}_m$, $m \in \{1, ..., M\}$ is removed if $\alpha \geqslant 0$:

$$\frac{Err^w_{null}(\mathcal{T}_m) - Err^w(\mathcal{T}_m)}{\rho(\mathcal{T}_m) - 1} < \alpha, \tag{1}$$

where $\rho(\mathcal{T}_m)$ is the number of terminal nodes in the sub-tree $\mathcal{T}_m$, $Err^w(\mathcal{T}_m)$ is the weighted misclassification error of the training samples within the sub-tree $\mathcal{T}_m$, and $Err^w_{null}(\mathcal{T}_m)$ is the weighted misclassification error for the sub-tree if it is terminated (concluded as a single terminal node). By sliding the value of α, we obtain a sequence of trees with shrinking sizes as α increases, although with growing misclassification errors. Two special cases are: $\alpha = 0$, which results in the original fully grown tree; and $\alpha = +\infty$, which results in a null tree. In practice, the value of α can be selected based on preferences of the tree size or cross-validation.

Web Appendix B: Tamoxifen treatment for breast cancer patients

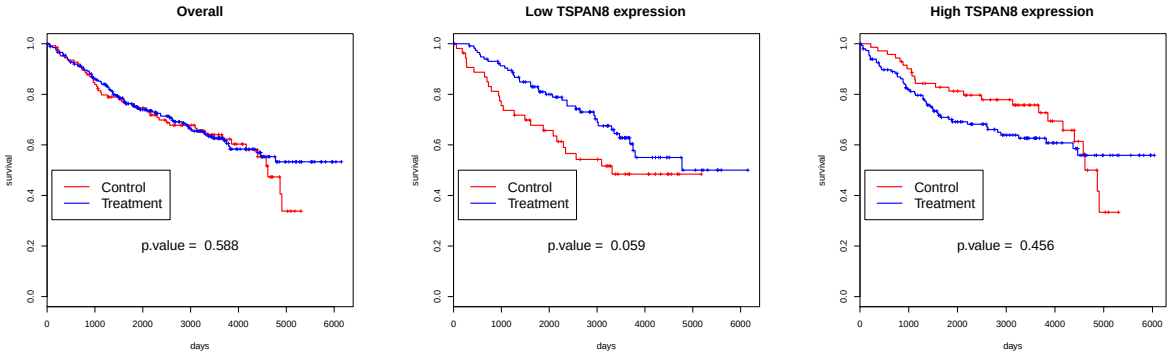
We apply our method to the cohort GSE6532 of data collected by Loi et al. (2007). The dataset consists of 277 patients who received Tamoxifen treatment and 137 patients in the control arm. Right censored survival outcomes (distant metastasis free survival time) are observed due to incomplete followup. Clinical information and a total of 44,928 gene expression measurements are available. In this analysis, we include three clinical variables:

age, grade, and size, and further include 500 genes that have the largest activity levels (marginal variance). Subjects with missing outcomes are removed from the analysis, which leads to our final dataset of 393 observations.

A main challenge in this analysis is that the treatment was not randomly assigned. Hence, we implement a simple extension of our proposed method to account for observational studies. Since $P(A|\mathbf{X})$ is unknown, we model this conditional probability in (9) or (5), and treat it as a plug-in estimator. This can be easily achieved by again using a tree-based method since it automatically provides the conditional probability estimation in a classification problem. The rest of the estimation procedure carries through naturally. Our method outputs a final model with two subgroups using gene expression of TSPAN8 (probe 203824_*at*). The protein encoded by this gene plays a role in the regulation of cell development, activation, growth and motility (Nazarenko et al., 2010), thus it has been found to be associated with a variety of cancers. In our final model, for the low TSPAN8 expression group ($\leqslant -0.063$), treatment is better than the control with a marginally significant log-rank test (p-value 0.059). For the high TSPAN8 group, no significant difference is detected (p-value 0.456). The Kaplan-Meier plots are provided in Web Figure 1.

Web Figure 1

Analysis of Tamoxifen dataset



This figure appears in color in the electronic version of this article.

References

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J. A., et al. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of Clinical Oncology* **25,** 1239–1246.

Nazarenko, I., Rana, S., Baumann, A., McAlear, J., Hellwig, A., Trendelenburg, M., Lochnit, G., Preissner, K. T., and Zöller, M. (2010). Cell surface tetraspanin tspan8 contributes to molecular pathways of exosome-induced endothelial cell activation. *Cancer Research* **70,** 1668–1678.