Survival Data Analysis (BIOS 7210)
Breheny

Final Project
Due: Monday, December 11

For the final project in this class, you will analyze a large, complex, time-to-event data set and write a paper on your conclusions. The data is available on the course website along with a detailed description of the variables and what they mean.

Analysis: The goal is to accurately model the time-to-event distribution, with the explanatory variables explaining as much variability as possible. In complex data sets, there are always a variety of interesting phenomena to explore: effects may be nonlinear, there may be interactions, proportional hazards might not hold, there may be strange outliers . . . I encourage you to be creative in your approach to modeling. Blindly following automatic stepwise procedures for adding and removing variables can prevent you from seeing important relationships in the data.

At the same time, be careful about overfitting. If an effect is fairly linear, don't feel compelled to include a spline just because you can. Furthermore, if your model is so complex that you no longer understand it, you should probably simplify it.

Thinking about whether a model makes biological/scientific sense is also important. Unless you know a lot about arrhythmias for some reason, you will probably need to read about some of the variables involved in this study to know whether the way in which you are modeling them is reasonable.

Finally, don't feel compelled to only include significant terms in your model. If a predictor *should* affect survival, but for whatever reason doesn't, that may be interesting to report as well.

Paper: The paper should have the same format as a regular scientific article:

- Introduction – Relevant biological/medical background to the problem being studied
- Methods – For our purposes, this should mainly describe the model-building process. How did you arrive at the final model you did, and why did you rule out other models? In particular, if you checked for things like interactions, nonlinear effects, but didn't include them in your model, or if you decided not to include certain predictors, explain this and justify your decisions.
- Results – This should provide summary and descriptive statistics as well as your final model(s). Plots and tables are typically very helpful here, although you must also describe in words what those plots and tables illustrate. Your results section should be the largest section, and should probably have subsections. For example, if you have a complicated term like an interaction your model, you might want to devote a subsection to explaining it.
- Discussion – A brief discussion of issues/limitations with the data and/or your model, along with your primary conclusions.

Altogether, the paper should be somewhere in the neighborhood of 10 pages, although your report may be somewhat longer or shorter depending on how you format things and how many plots you include.

Predictions: Since the goal of this project is to identify at-risk patients, I thought it would be interesting (fun?) to add a prediction component to the project. In addition to the 946 observations in the data set, there are 500 subjects for whom I have provided you the predictors, but not the outcome. Your job is to predict whether the patient will have experienced an ICD shock during the first 3 years (36 months) of follow-up.

Specifically, for each patient i in the prediction set, you are to provide $\hat{S}(36|\mathbf{x}_i)$, your estimated probability that the patient will remain shock-free at 36 months. To evaluate the accuracy of these predictions, I will calculate the Kullback-Leibler loss against y_i, the actual shock status of the patient ($y_i = 1$ if no shock by 3 years, and $y_i = 0$ if shock before 3 years):

$$\sum_i -\{y_i \log \hat{S}(36|\mathbf{x}_i) + (1 - y_i) \log(1 - \hat{S}(36|\mathbf{x}_i))\}$$

For the prediction component, create a numeric vector of length 500 with your survival predictions. Save the file with

```
saveRDS(pred, file='PatrickBreheny.rds')
```

(where `pred` is the name of your prediction vector and you should replace `PatrickBreheny` with your actual name) and e-mail the file to me. I will calculate everyone's predictive accuracy and send out an e-mail letting you know how you did, as well as an e-mail to the class announcing who the overall winner was.