Introduction to 2018 Statistics Methods Forum Data Challenge

Eric Polley

June 27th, 2018

Introduction

The focus this year is the estimation and evaluation of a prognostic risk score with a right censored outcome of interest.

The dataset is from a clinical trial in Non-Small Cell lung cancer with N=398 patients. The patients were randomly split into training (N=300) and a blinded test set (N=98).

Details for the Data Challenge are available on Github:
https://github.com/ecpolley/Data_Challenge_2018

Outline

The data challenge will use the next 2 Statistical Methods Forum meetings

- June 27th: Introduction to the data challenge
- July 25th: Group Discussion and Q&A
- August 20th 5:00pm local: Team submission deadline
- August 22nd: Final Results presentation

Team Science

- Participants are encouraged to work in teams ($N \in (1, 2, \ldots, 10)$)
- Opportunity to learn from each other and work with people outside usual team
- Data is publicly available, so is available outside Mayo
- If you would like help forming a team, email Eric Polley or Kristin Mara
- Teams are responsible for creating a team name, and may submit up to 3 estimates, with the last submission being the official one
- If you are participating, please let us know in case we have any Data Challenge announcements

Overview of Dataset

- Two datasets will be provided for the training and test sets
- Clinical dataset with baseline variables and outcome
- Each patient had a baseline Lung CT scan, single slice compiled in an (512, 512, N) Array

Overview of Dataset

```
# link to data on GitHub page if not available
if(file.exists("Training_clinical.csv")) {
  dat <- read.csv("Training_clinical.csv")
} else {
  urlfile <- "https://raw.githubusercontent.com/ecpolley/
    Data_Challenge_2018/master/Training_clinical.csv"
  download.file(urlfile, destfile = "Training_clinical.csv'
  dat <- read.csv("Data.csv")
}
dim(dat)
```

```
## [1] 300  12
```

setup

```r
library(arsenal)
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

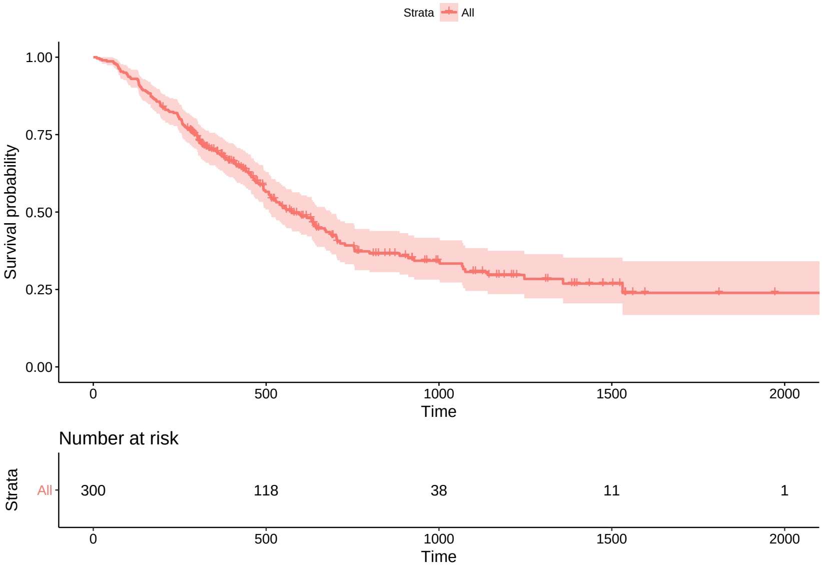
```
## Loading required package: ggpubr
```

```
## Loading required package: magrittr
```

Overview of Dataset

	Overall (N=300)
age	
Mean (SD)	68.261 (10.250)
Range	33.685 - 91.704
as.factor(Clinical.T.Stage)	
1	57 (19.0%)
2	111 (37.0%)
3	43 (14.3%)
4	89 (29.7%)
as.factor(Clinical.N.Stage)	
0	123 (41.0%)
1	18 (6.0%)
2	103 (34.3%)
3	54 (18.0%)
4	2 (0.7%)
Clinical.M.Stage	
Mean (SD)	0.000 (0.000)
Range	0.000 - 0.000
Overall.Stage	
I	60 (20.0%)
II	32 (10.7%)
IIIa	84 (28.0%)
IIIb	124 (41.3%)
Histology	
N-Miss	24
adenocarcinoma	36 (13.0%)
large cell	85 (30.8%)
nos	42 (15.2%)
squamous cell carcinoma	113 (40.9%)
gender	
female	95 (31.7%)
male	205 (68.3%)

Overview of Dataset



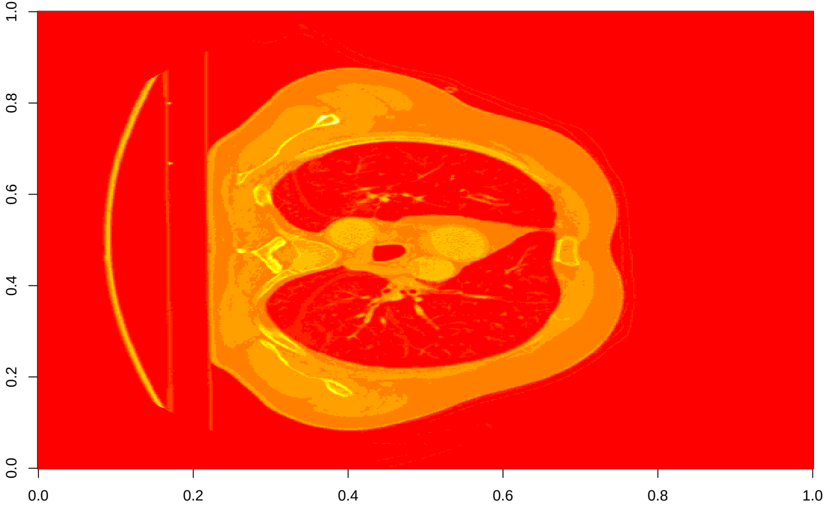
Overview of Dataset

```r
load("DataChallengeDataTrain_array.RData")
dim(IMAGES_array_train)
```

```
## [1] 512 512 300
```

Overview of Dataset

```r
image(IMAGES_array_train[, , 4])
```



Goal

- The primary goal is to develop a prognostic risk score
- How to evaluate on the held out set?

Evaluation

Primary goal:

- Each each patient in test set, provide predicted risk score
- Evaluate discrimination by estimating concordance with observed survival times
- Each team can email me (Polley.Eric@Mayo.edu) with text file including patient ID and predicted risk score

Evaluation

Secondary goal:

- Evaluate calibration of predicted probability of survival at specific time points
- For each patient in test set, provide predicted probability of survival at 1, 2, and 3 years post treatment
- 365, 730, and 1095 days
- For each time point, split test data into quintiles based on predicted probability. Compare to Kaplan-Meier estimate.

Questions?