Detection of Prostate Cancer

November 14, 2018

Abstract

Prostate cancer is one of the most common cancers among men. Luckily, prostate cancer is treatable when detected early enough. The Cancer Center at Ohio State University wants to know if baseline exam measurements can predict whether a tumor has penetrated the prostatic capsule indicting there is indeed a tumor present. The Prostatic Specific Antigen value (PSA), results of digital exam (dpros), and Gleason score has the ability of predicting whether a patient has a cancerous tumor penetrating the prostatic capsule. In the end, this model does not have any interaction terms even though they were considered within the scope of the analysis.

1 Introduction

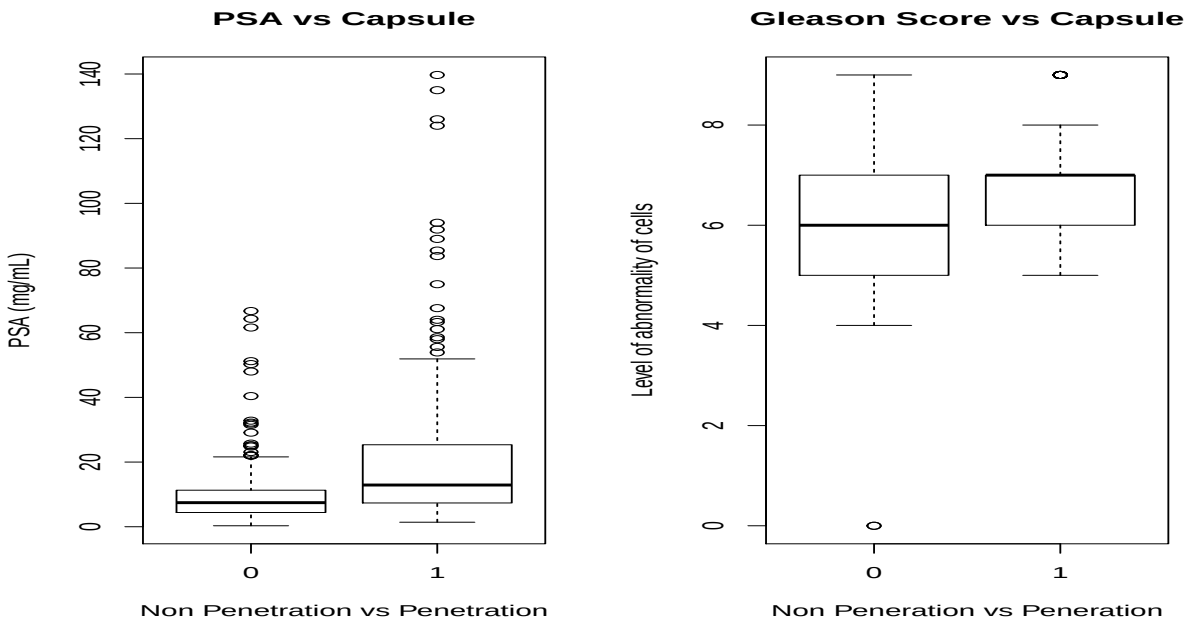
Since prostate cancer is one of the most common cancer among men, a study was conducted that the Comprehensive Cancer Center of Ohio State whether it was possible to determine if a tumor has penetrated the prostatic capsule. This is good news for a majority of men if this analysis can accurately determine what causes prostate cancer which occurs about 1 out of 7 men, but is only 1 in 39 men will die due to this cancer. Several factors contribute to cancer tumors penetrating the prostatic capsule which include: age of subject, race, results of digital exam, detection of capsular involvement, Prostatic Specific Antigen value, and total Gleason score. It appears that the Prostatic Specific Antigen value (PSA), results of digital exam, and Gleason score is able predicting whether a patient has a cancerous tumor penetrating the prostatic capsule.

2 Methods

There was a cancer study on 380 male patients of either white or black race. Patients 1162, 1186, 1392 were excluded because those patients had at least one "na" value listed. For those who do not know about prostate cancer, many of the variables will seem unfamiliar. PSA is a measure of protein produced by prostate gland cells and is measured in mg/mL. Elevated levels of PSA may suggest prostate cancer and is used as a screening test. The Gleason score is a scale from 1 to 10 measuring the abnormality of cells. Larger values of Gleason score suggest a higher risk of cancer. Race has two factors, whether the patient is black or white. The variable capsule indicates whether the tumor penetrated the prostatic capsule. dpros are the results of the digital rectal exam which can have no nodule, unilobar nodule, bilobar nodule. dcaps is the detection of capsular involvement. There was 153 of the 380 subjects who had a cancer that penetrated the capsule.

This analysis will not have any transformation, but will consider interaction terms. The analysis will be conducted using R/RStudio.

Figure 1: Boxplot of PSA/Gleason vs Capsule Penetration



3 Results

3.1 Exploratory Data Analysis

In figure 1, it shows two things, the amount of PSA when the capsule is penetrated and the probability of having the capsule penetrated given the level of abnormality in the cells. This did show some interesting results. PSA is higher when the capsule is penetrated and the Gleason score is generally higher when there is penetration. These two graphs shows that the doctors and scientists Ohio State were onto something. Also, a simple Chi-squared t-test located in the appendix would show that dpros has a significant association with the capsule being punctured.

3.2 Model Fitting/Inferences

By using the stepAIC on all possible models containing interaction terms, the model of with the least of amount of penalties still had non significant terms. Manually, the terms that were not significant by p-values were dropped one-by-one out of the model. The model contained three predictor variables there were not highly correlated with each other shown in figure 2 in the appendix. just at a glance, the correlation plot also shows that the Gleason score is moderately correlated to a tumor penetration of the prostatic capsule.

Interpreting an odds ratio is is not too difficult. Any odds ratio at 1 does not influence the response. Odds ratio that are between 0 and 1 are (1 - odds ratio) *100 % less odds of having the capsule penetrated by a tumor, but odds ratio greater than 1, then the patient has (odds ratio - 1) *100 % more odds of having the capsule penetrated by a tumor. For example, PSA has a coefficient of .03 and has 1.03 odds ratio which means that with 10 mg/mL increase in PSA while all other variables are held constant, then that patient has 30% increase in odds of having the capsule penetrated by a tumor according to table 1. The rest of the odds ratio can be interpreted in a similar manner, but be cautious when interpreting the intercept or any of the dpros predictors. The intercept should be seen as no penetration or dpros1, whereas dpros2 and dpros 3 are directional. In addition, there are confidence intervals associated with their respective odds ratio.

With this model, a patient's odds of a tumor penetrating their prostatic capsule can be determined.

Table 1: Odds Ratio with 95% CI on Odds Ratio

	Coefficient	Odds Ratio	SE	p-value	95% CI on OR
intercept	-8.14	0.00	1.06	0.0000	(0.00 , 0.00)
dpros2	0.77	2.17	0.36	0.0300	(1.09 , 4.43)
dpros3	1.55	4.73	0.37	0.0000	(2.32 , 10.00)
dpros4	1.43	4.18	0.45	0.0015	(1.75 , 10.20)
psa	0.03	1.03	0.01	0.0036	(1.01 , 1.05)
gleason	0.99	2.71	0.16	0.0000	(2.00 , 3.76)

Table 2: Predictions on Simulated values

Odds Ratio	dpros	PSA	Gleason
0.76	unilobar nodule (right)	14.1	7
1.41	unilobar nodule (left)	30	8
2.93	bilobar nodule	25	9

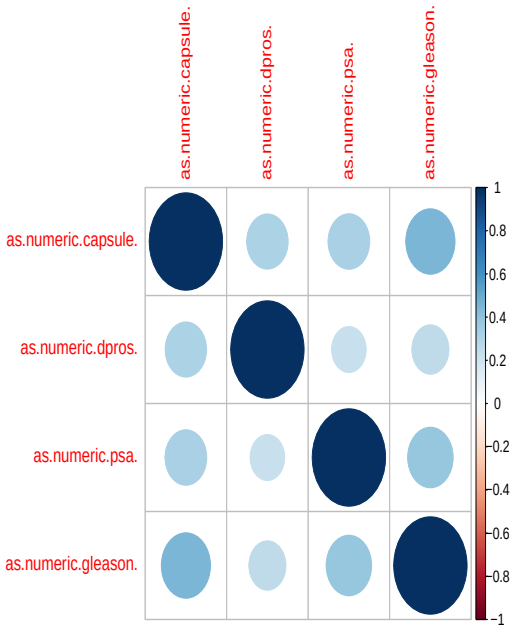
Based on table 2, there are several values that were simulated to see a patient's odds of whether a tumor has penetrated their prostatic capsule. For an example, take a patient that has a unilobar nodule on his right side, with PSA of 14.1 and a total Gleason score of 7, then that patient has (1 - .76) * 100 % or 24% less likely odds of having a tumor protruding into their capsule.

4 Conclusions

The analysis has shown that total Gleason score, Prostatic Specific Antigen value, and results of digital rectal exam are the best variables that can predict the probability a man will have a tumor penetrating the prostatic capsule. Refer to table 2 to see some simulated patients and their odds of having a tumor penetrate the prostate capsule. There were several limitations of this study. Age did not play a major factor in the model since the study focused on older men with the youngest man being 47 in the study. If the study included younger men from the ages of 18-35, one can speculate that age will play a huge factor if those young men has prostate cancer. In addition, a table showing race would say that only 36 patients were black. That number may or may not be enough to figure out whether this model is good for black patients, but a simple power calculation would suffice. An important question for doctors and scientists is whether prostate cancer can be predicted before a cancerous tumor has penetrated the prostatic capsule. If prostate cancer can be detected even before it has punctured the capsule, then the rate of survival can be raised even higher than the current rate.

Figure 2: Correlation plot among Covariates and Response

```
## corrplot 0.84 loaded
```



Appendix A: Auxiliary Graphics and Tables

```
## age
## 47 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 7
##  1  2  3  2  4 10  5  7 10 10 13  9 13 20 11 17 24 23 16 28 23 28 24 1
## 74 75 76 77 78 79
## 13 13  9  4  5  2
## race
##   1   2
## 341  36
## capsule
##   0   1
## 226 151
```

```
##        race
## capsule   1   2
##       0 204  22
##       1 137  14
##        dpros
## capsule  1  2  3  4
##       0 80 83 45 18
##       1 19 48 50 34
##
## Call:
## glm(formula = capsule ~ dpros + psa + gleason, family = binomial(link =
##     data = Prostate)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3555  -0.7582  -0.4434   0.9102   2.4625
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.144845   1.056459  -7.710 1.26e-14 ***
## dpros2       0.772554   0.355944   2.170  0.02997 *
## dpros3       1.553114   0.371366   4.182 2.89e-05 ***
## dpros4       1.429280   0.449133   3.182  0.00146 **
## psa          0.027336   0.009401   2.908  0.00364 **
## gleason      0.995310   0.161133   6.177 6.54e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 507.61  on 376  degrees of freedom
## Residual deviance: 381.22  on 371  degrees of freedom
## AIC: 393.22
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
##
## Cohen's d
##
## d estimate: 1.057012 (large)
## 95 percent confidence interval:
##      inf      sup
## 0.904368 1.209656
##
## Cohen's d
##
## d estimate: 7.060166 (large)
## 95 percent confidence interval:
##      inf      sup
## 6.675678 7.444655
##
## Cohen's d
##
## d estimate: 2.368822 (large)
## 95 percent confidence interval:
##      inf      sup
## 2.182314 2.555330
```

```
##
##  Pearson's Chi-squared test
##
## data:  dpros and capsule
## X-squared = 38.736, df = 3, p-value = 1.974e-08
```

Appendix B: R Code

```
# Read-in data
setwd("~/SchoolWork/Stat696/Prostate")
Prostate = read.table("prostate.txt", header=T)
dim(Prostate); names(Prostate)
#[1] 380 8
sum(is.na(Prostate)) # 3 in race
Prostate <- na.omit(Prostate); dim(Prostate) # 377 8
Prostate$dcaps = as.factor(Prostate$dcaps)
Prostate$dpros = as.factor(Prostate$dpros)
Prostate$race = as.factor(Prostate$race)
Prostate = Prostate[, -1]
attach(Prostate)
head(Prostate)
summary(Prostate)
library(MASS)
library(effsize)
library(VIF)
library(xtable)
library(corrplot)

# A reminder: in this project, we will consider interactions, but not consider transformations.
# a) binary response, capsule: consider a simple table of counts
table(capsule)

# b) binary response against categorical covariates: consider tables
#    (can perhaps consider side-by-side bar charts, though I think contingency tables are sufficient)
table(capsule, race)
table(capsule, dpros)
table(capsule, age)
# c) binary response against continuous covariates: consider summary tables and box-plots
boxplot(psa ~ capsule, main = "PSA vs Capsule", xlab = "Non Penetration vs Penetration", ylab = "PSA (mg/mL)")
boxplot(age ~ capsule, main = "Age vs Capsule", xlab = "Non Peneration vs Peneration", ylab = "Age")
boxplot(gleason~capsule, main = "Gleason Score vs Capsule", ylab = "Level of abnormality of cells", xlab = "Non Peneration vs Peneration")

# d) standardized mean difference: evaluation of relationships between covariates and the response
cohen.d(age, capsule)
cohen.d(psa, capsule)
cohen.d(gleason, capsule)
cohen.d(as.numeric(dpros), capsule)

## Task 2: Model building
# Logistic regression models are fit using the glm function using the logit link:
#   e.g., glm(capsule~psa+gleason, family=binomial(link=logit), data=Prostate)

fit = glm(capsule~ .*., family = binomial(link=logit), data = Prostate)
fit1 = glm(capsule ~  dpros + psa + gleason, family = binomial(link = logit), data = Prostate)

provars = data.frame(as.numeric(capsule), as.numeric(dpros), as.numeric(psa), as.numeric(gleason))
cpv = cor(provars)
corrplot(cpv)


# a) Stepwise model selection: include interactions, consider stepAIC for first pass
stepAIC(fit)
stepAIC(fit1)
```

```r
# b) Parsimonious model: perform backward selection via p-values,
#      identify a simpler model by being strict with interaction terms.
#      Is it that much worse than best stepwise model?
summary(fit)
summary(fit1)
vif(fit1)
## Task 3: Model evaluation
# Prostate_diagnostics_ClassVersion.R presents sample code for a given model via explanatory variable patterns (EVPs).
# The components include:
# a) Residual plots: use the examine.logistic.reg function provided
# b) Outlier detection: evaluate EVPs for potential outlying data points
# c) HL test of overall fit: use the HLtest.R function provided; see PK_diagnostics_ClassVerion.R

# Residual plots
# Load-in required functions
one.fourth.root=function(x){
  x^0.25
}
source("examine.logistic.reg.R")

# Consider model of PSA, Gleason score, and Detection of capsular involvement
dat.glm <- glm(capsule ~ psa+gleason+dcaps, family = binomial, data = Prostate)
dat.mf <- model.frame(dat.glm)
## Covariate pattern: too many EVPs!
w <- aggregate(formula = capsule ~ psa+gleason+dcaps, data = Prostate, FUN = sum)
n <- aggregate(formula = capsule ~ psa+gleason+dcaps, data = Prostate, FUN = length)
w.n <- data.frame(w, trials = n$capsule, prop = round(w$capsule/n$capsule,2))
dim(w.n)
#[1] 301 6

# Create EVPs by binning continuous covariates
g = 5 # number of categories
psa_interval = cut(psa, quantile(psa, 0:g/g), include.lowest = TRUE)  # Creates factor with levels 1,2,...,g
levels(psa_interval)

# Diagnostic plots
w <- aggregate(formula = capsule ~ psa_interval+gleason+dcaps, data = Prostate, FUN = sum)
n <- aggregate(formula = capsule ~ psa_interval+gleason+dcaps, data = Prostate, FUN = length)
w.n <- data.frame(w, trials = n$capsule, prop = round(w$capsule/n$capsule,2))
mod.prelim1 <- glm(formula = capsule/trials ~ psa_interval+gleason+dcaps,
                   family = binomial(link = logit), data = w.n, weights = trials)
save1 = examine.logistic.reg(mod.prelim1, identify.points=T, scale.n=one.fourth.root, scale.cookd=sqrt)

# Evaluation of EVPs for potential outlying sets of points
w.n.diag1 = data.frame(w.n, pi.hat=round(save1$pi.hat, 2), std.res=round(save1$stand.resid, 2),
                       cookd=round(save1$cookd, 2), h=round(save1$h, 2))
p = length(mod.prelim1$coef) # number of parameters in model (# coefficients)
ck.out = abs(w.n.diag1$std.res)>2 | w.n.diag1$cookd>4/nrow(w.n) | w.n.diag1$h > 3*p/nrow(w.n)
extract.EVPs = w.n.diag1[ck.out, ]
extract.EVPs

# Note: EVPs purely for diagnostics, akin to histogram binning to assess
# distribution shape.  The analysis does not use this EVP binning.

betahat = formatC(signif(fit1$coeff, digits = 3), digits = 2, format = "f", flag = "#")
OR = formatC(signif(exp(fit1$coeff), digit = 3), digits = 2, format = "f", flag = "#")
SE = formatC(signif(summary(fit1)$coeff[,2], digits = 3), digits = 2, format = "f", flag = "#")
```

```
cibounds = formatC(signif(exp(confint(fit1)), digits = 3), digits = 2, format = "f", flag =
    "#")
pval = formatC(signif(summary(fit1)$ coeff[,4], digits = 4), format = "f", flag = "#")

x = cbind(betahat, OR, SE, pval, matrix(paste("(", cibounds[,1], "," , cibounds[,2], ")")))
colnames(x) = cbind("Coefficient", "Odds Ratio", "SE", "p-value", "95% CI on OR")
rownames(x) = cbind("intercept", "dpros2", "dpros3", "dpros4", "psa", "gleason")
inftable = xtable(x)
align(inftable) = "|l|ccccc|"
print(inftable)

###EDA###
table(race)
table(capsule)
table(capsule,race)
table(capsule,dcaps)
table(capsule,dpros)

prop.table(table(capsule))
prop.table(table(capsule,race))
prop.table(table(capsule,dcaps))
prop.table(table(capsule,dpros))


boxplot(age~capsule, data = Prostate)
boxplot(psa~capsule, data = Prostate)
boxplot(gleason~capsule, data = Prostate)
plot(dpros, capsule, data = Prostate)

cohen.d(age, capsule)
cohen.d(psa, capsule)
cohen.d(capsule, race)
cohen.d(gleason, capsule)

t.test(age[capsule==0], age[capsule==1])
t.test(psa[capsule==0], psa[capsule==1])
t.test(gleason[capsule==0], gleason[capsule==1])
table(age)
table(race)
## prediction
new.data = data.frame(dpros = as.factor(3), psa = 14.1, gleason = 7)
new.data2 = data.frame(dpros = as.factor(2), psa = 30, gleason = 8)
new.data3 = data.frame(dpros = as.factor(4), psa = 25, gleason = 9)
predict(fit1, new.data, interval = "prediction")
predict(fit1, new.data2, interval = "prediction")
predict(fit1, new.data3, interval = "prediction")
chisq.test(x = dpros, y = capsule)
```

Listing 1: Prostate Cancer