Effects of Demographic Factors and Genes on Biochemical Recurrence Prostate Cancer Among Patients

Yiyan(Shelley) NI

Background

Biochemical recurrence(BCR) continues to occur in a large proportion of prostate cancer patients. Yet, established clinical predictors(PSA, Gleason score) often provide poor prognosis. It's noticeable that more research is now focusing on the effect of genes such as integrated genomics and gene signatures on prostate cancer BCR.

Objective

To investigate whether the target genes as well as demographic factors have effects on biochemical recurrence(BCR) free time among prostate cancer patients in three public datasets(Taylor, Cambridge and Stockholm). Also, to test if BCR free times are significantly different among three datasets. In addition, comparing the results with the previous research from a shiny app.

Introduction

Prostate cancer is the most commonly diagnosed cancer among Canadian men. It's the 3rd leading cause of death from cancer in men in Canada. In 2017, it's estimated that 21,300 men were diagnosed with prostate cancer. This represents 21% of all new cancer cases in men in 2017. On average, 58 Canadian men were diagnosed with prostate cancer and 11 Canadian men died from prostate cancer every day.

A genetic contribution to prostate cancer risk has been documented. Mutations in BRCA1 and BRCA2, important risk factors for ovarian cancer and breast cancer in women, have also been implicated in prostate cancer.[1] Other linked genes include the Hereditary Prostate cancer gene 1 (HPC1), the androgen receptor, and the vitamin D receptor.[2] TMPRSS2-ETS gene family fusion, specifically TMPRSS2-ERG or TMPRSS2-ETV1/4 promotes cancer cell growth.[3] Also, Two large genome-wide association studies linking single-nucleotide polymorphisms (SNPs) to prostate cancer were published in 2008.[4][5] These studies identified several SNPs which substantially affect the risk of prostate cancer. In addition, demographic factors such as obesity, age and high blood pressure are also found to be associated to prostate cancer.

Section 1 Univariate Analysis

After combining and cleaning these three datasets(Taylor, Cambridge and Stockholm), we conduct a univariate analysis based on three datasets(see Table 1).

Table 1. Patient Baseline Characteristics

Characteristics	ALL (N=343)	Cam (N=111)	Stoch (N=92)	Taylor (N=140)	p.overall	N
Age, mean (SD), y	17.6 (6.3)	17.6 (6.3)	. (.)	. (.)	.	111
Time, mean (SD), m	51.5 (29.7)	35.6 (17.1)	78.9 (21.0)	46.0 (30.3)	<0.001	343

Table 1. Patient Baseline Characteristics (Continued)

Characteristics	ALL (N=343)	Cam (N=111)	Stoch (N=92)	Taylor (N=140)	p.overall	N
PSA, mean (SD),ng/ml	9.7 (9.3)	8.6 (3.7)	11.0 (13.2)	. (.)	0.063	200
Event, No. (%)					<0.001	343
0	243 (70.8%)	92 (82.9%)	47 (51.1%)	104 (74.3%)		
1	100 (29.2%)	19 (17.1%)	45 (48.9%)	36 (25.7%)		
Stage, No. (%)					<0.001	341
1c	181 (53.1%)	61 (55.0%)	41 (45.1%)	79 (56.8%)		
2	17 (5.0%)	17 (15.3%)	0 (0.0%)	0 (0.0%)		
2a	62 (18.2%)	5 (4.5%)	34 (37.4%)	23 (16.5%)		
2b	31 (9.1%)	8 (7.2%)	4 (4.4%)	19 (13.7%)		
2c	17 (5.0%)	4 (3.6%)	1 (1.1%)	12 (8.6%)		
3	6 (1.8%)	4 (3.6%)	0 (0.0%)	2 (1.4%)		
3a	23 (6.7%)	10 (9.0%)	9 (9.9%)	4 (2.9%)		
3b	2 (0.6%)	2 (1.8%)	0 (0.0%)	0 (0.0%)		
Tx	2 (0.6%)	0 (0.0%)	2 (2.2%)	0 (0.0%)		
iCluster, No. (%)					<0.001	329
1	59 (17.9%)	23 (22.1%)	24 (28.2%)	12 (8.6%)		
2	81 (24.6%)	24 (23.1%)	23 (27.1%)	34 (24.3%)		
3	34 (10.3%)	9 (8.7%)	7 (8.2%)	18 (12.9%)		
4	49 (14.9%)	19 (18.3%)	5 (5.9%)	25 (17.9%)		

Table 1. Patient Baseline Characteristics (Continued)

Characteristics	ALL (N=343)	Cam (N=111)	Stoch (N=92)	Taylor (N=140)	p.overall	N
5	63 (19.1%)	29 (27.9%)	26 (30.6%)	8 (5.7%)		
6	21 (6.4%)	0 (0.0%)	0 (0.0%)	21 (15.0%)		
flat	22 (6.7%)	0 (0.0%)	0 (0.0%)	22 (15.7%)		
Gleason, No. (%)					0.001	339
10	1 (0.3%)	0 (0.0%)	1 (1.1%)	0 (0.0%)		
6	76 (22.4%)	17 (15.3%)	18 (20.0%)	41 (29.7%)		
7	216 (63.7%)	85 (76.6%)	55 (61.1%)	76 (55.1%)		
8	24 (7.1%)	8 (7.2%)	5 (5.6%)	11 (8.0%)		
9	20 (5.9%)	1 (0.9%)	9 (10.0%)	10 (7.2%)		
5	2 (0.6%)	0 (0.0%)	2 (2.2%)	0 (0.0%)		
ECE, No. (%)					0.002	201
N	82 (40.8%)	34 (30.6%)	48 (53.3%)	0 (.%)		
Y	119 (59.2%)	77 (69.4%)	42 (46.7%)	0 (.%)		
SLC2A1, mean (SD)	7.1 (2.4)	8.5 (0.3)	3.1 (0.1)	8.6 (0.2)	0.000	343
SLC2A2, mean (SD)	5.0 (1.5)	6.0 (0.1)	2.5 (<0.1)	5.9 (0.4)	<0.001	343
SLC2A3, mean (SD)	7.4 (2.7)	9.4 (0.8)	3.1 (0.1)	8.7 (0.8)	<0.001	343
SLC2A4, mean (SD)	7.5 (3.7)	6.4 (0.1)	2.6 (<0.1)	11.7 (0.1)	0.000	343
SLC2A5, mean (SD)	6.2 (2.1)	7.7 (0.8)	2.8 (0.1)	7.3 (0.4)	<0.001	343

Table 1. Patient Baseline Characteristics (Continued)

Characteristics	ALL (N=343)	Cam (N=111)	Stoch (N=92)	Taylor (N=140)	p.overall	N
SLC2A6, mean (SD)	6.2 (2.2)	6.6 (0.3)	2.7 (0.1)	8.1 (0.3)	0.000	343
SLC2A7, mean (SD)	5.6 (1.9)	6.1 (0.1)	2.5 (<0.1)	7.2 (0.3)	0.000	343
SLC2A8, mean (SD)	7.1 (2.4)	9.1 (0.3)	3.2 (0.1)	8.1 (0.2)	0.000	343
SLC2A9, mean (SD)	5.8 (1.9)	6.5 (0.2)	2.7 (<0.1)	7.3 (0.3)	0.000	343
SLC2A10, mean (SD)	8.1 (2.9)	9.8 (0.4)	3.3 (0.1)	9.9 (0.3)	0.000	343
SLC2A11, mean (SD)	6.1 (2.1)	6.6 (0.2)	2.7 (<0.1)	7.9 (0.2)	0.000	343
SLC2A12, mean (SD)	7.2 (2.5)	9.1 (0.7)	3.3 (0.1)	8.4 (0.8)	<0.001	343
SLC2A13, mean (SD)	5.5 (1.9)	6.1 (0.1)	2.6 (<0.1)	7.1 (0.2)	0.000	343
SLC2A14, mean (SD)	6.2 (2.4)	6.3 (0.1)	2.6 (<0.1)	8.6 (0.6)	<0.001	343
HK1, mean (SD)	8.2 (3.1)	11.4 (0.3)	3.5 (<0.1)	8.8 (0.3)	0.000	343
HK2, mean (SD)	6.9 (2.5)	7.7 (0.4)	2.9 (0.1)	8.9 (0.3)	<0.001	343
HK3, mean (SD)	5.8 (2.0)	6.6 (0.2)	2.6 (<0.1)	7.3 (0.4)	<0.001	343
CHGA, mean (SD)	6.1 (2.1)	7.6 (0.9)	2.8 (0.2)	7.1 (0.6)	<0.001	343
NCAM1, mean (SD)	5.9 (2.0)	6.8 (0.3)	2.6 (<0.1)	7.3 (0.4)	<0.001	343
GCK, mean (SD)	5.8 (2.0)	6.4 (0.1)	2.6 (<0.1)	7.4 (0.4)	<0.001	343
FOLH1, mean (SD)	7.7 (3.0)	8.7 (1.2)	3.2 (0.2)	9.9 (1.2)	<0.001	343
SYP, mean (SD)	5.7 (1.9)	6.3 (0.4)	2.6 (0.1)	7.2 (0.4)	<0.001	343
ENO2, mean (SD)	6.1 (2.1)	7.3 (0.5)	2.7 (0.1)	7.4 (0.3)	<0.001	343

Some statistics and analysis can be noticed from Table 1 that

- There are 343 patients in total. 111 from Cam dataset, 92 from Stoch dataset and 140 from Taylor dataset.
- Factor—"Age" is unavailable in both Stoch and Taylor datasets. Factors—"PSA" and "ECE" are unavailable in Taylor dataset.
- There are 2 missing values in "Stage", 14 Missing values in "iCluster" and 4 missing values in "Gleason". Also, we assume the deleted missing values in "Time" and "Event" are random distributed.
- All the variables have a significant p-value(<0.05) among these three datasets except "PSA".

From Table 1, it's also clear that the demographic factors include

- Age: patient's age
- Time: patient's biochemical recurrence(BRC) free time
- PSA: patient's PSA(prostate-specific antigen) level
- Event: two levels, 0 (right censoring), 1 (relapse)
- Stage: it measures how far the cancer has spread. Nine levels.
- iCluster: seven levels.
- Gleason: Gleason Score. Six levels.
- ECE: extra-capsular extension. Two levels, Yes and No.

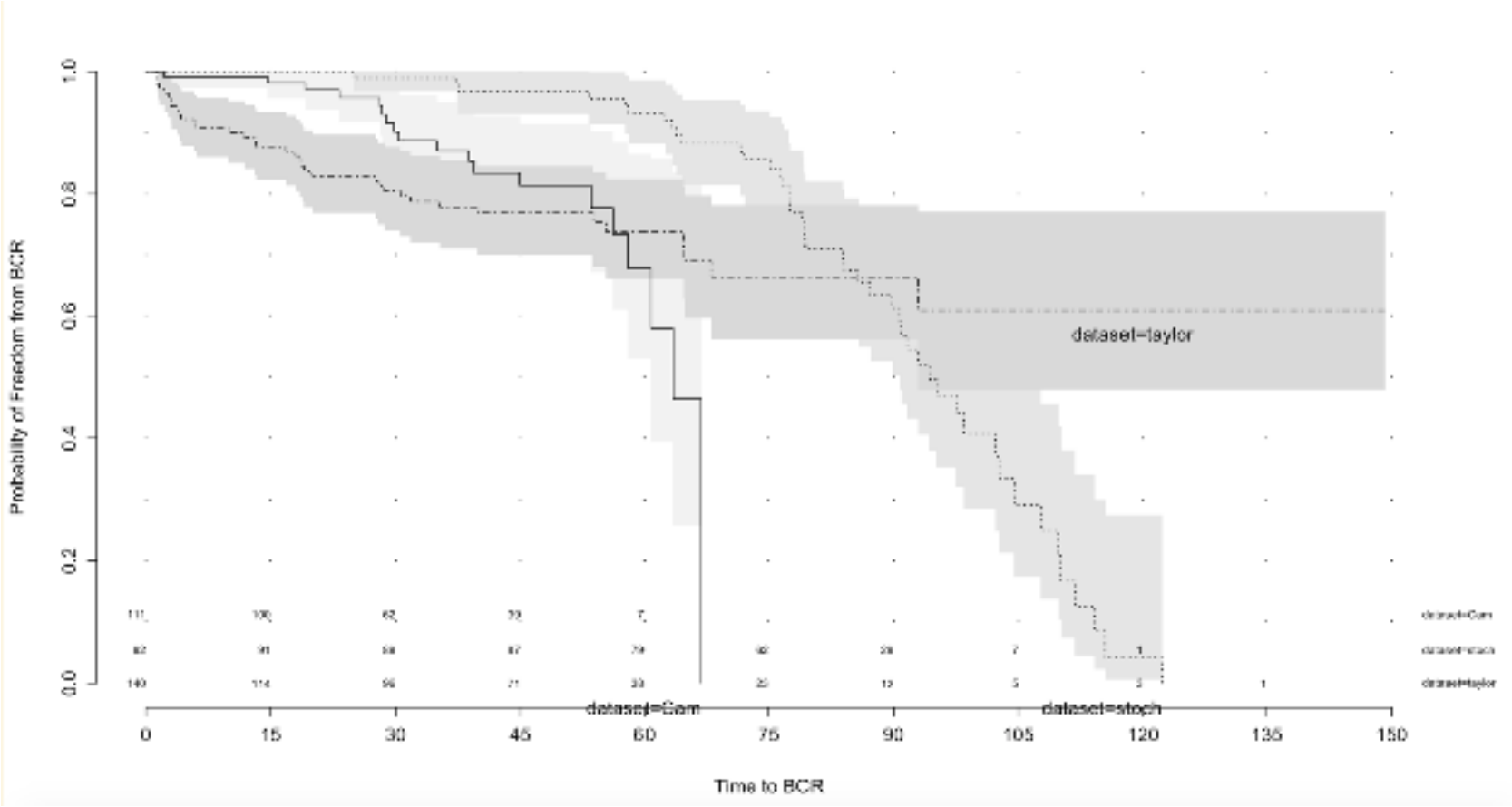
And the 23 target genes include:
SLC2A1-14, HK1-3, GCK, OAX1, ENO2, NCAM1, SYP, CHGA, FOLH1, KLK3

Section 2 KM Curves

2.1 KM curves among three datasets (Graph 1)

It's clear from Graph 1 that patients in dataset—Stoch have the highest survival rate before BCR free time turns 90 months. The patients in dataset—Cam have the second highest survival rate until BCR free time turns 60 months, after which the survival rate plumps to 0 when the BCR free time is about 67 months. The patients in dataset—Taylor have the

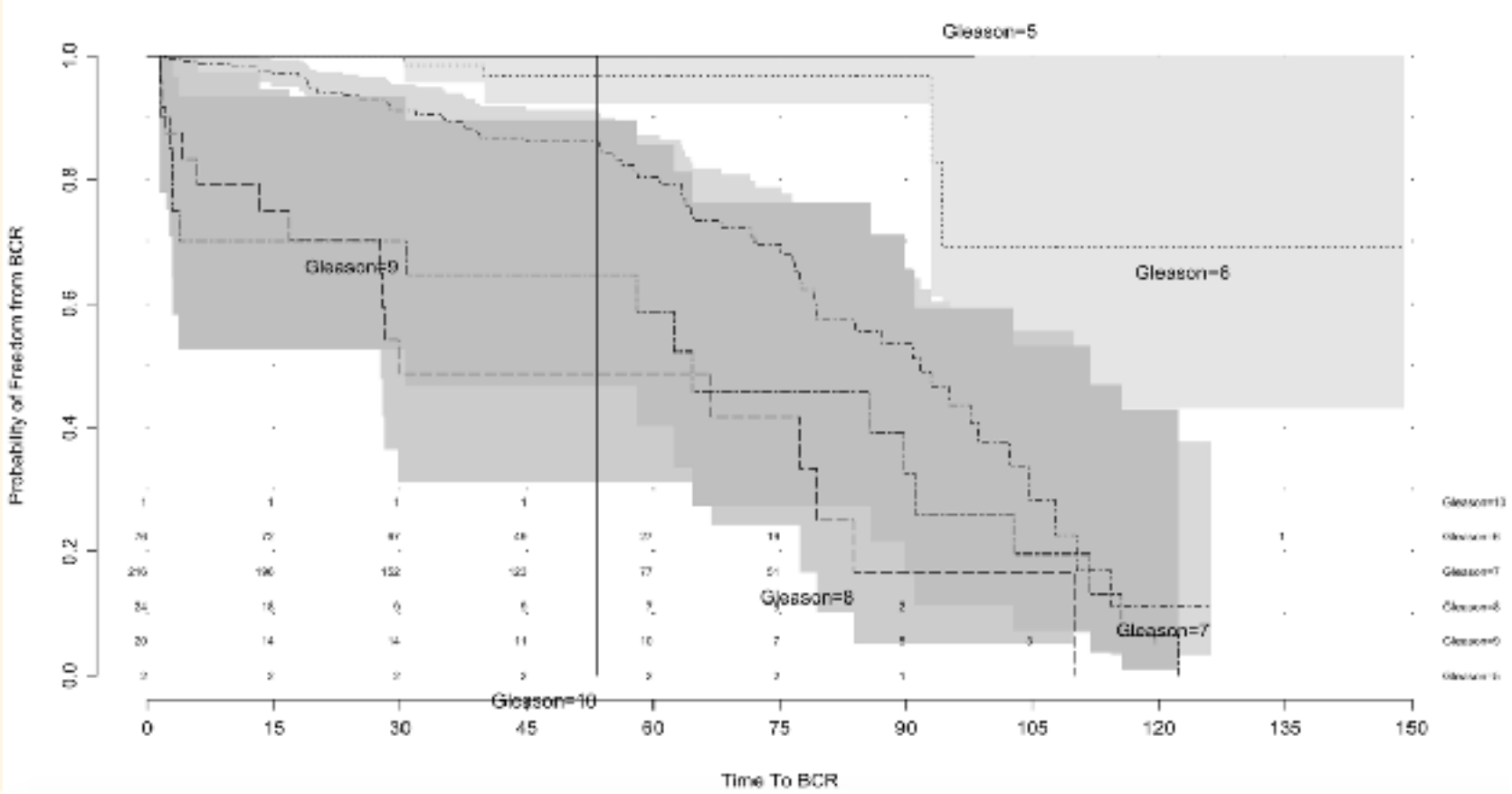
lowest survival time at the beginning and becomes the second in the end. Even though the curves have 95% CIs, it's still hard to tell if they are significantly different among three datasets.



Graph 1 KM curves among three datasets

2.2 KM curves among different Gleason Scores(Graph 2)

From Graph 2, it's uneasy to tell if the survival rates of patients having different Gleason Scores are significantly different. Also, not enough observations for Gleason Scores are 5 and 10.



Graph 2 KM curves among Gleason Scores

Section 3 Logrank Test among three datasets

3.1 Logrank test

It's uneasy to see whether the statistically difference exists among three datasets from Graph 1 since the KM curves cross each other and so on. So, we now construct a logrank test since we want to investigate if the probabilities of freedom from BCR among three datasets are significantly different.

From the R output 1, we learn that p= 0.3 so we can't reject the survival rates are the same among three datasets based on the Logrank Test.

R output 1:

survdiff(formula = Surv(Time, Event) ~ dataset, data = datacamtaylorstock,
rho = 0)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
dataset=Cam	111	19	14.6	1.33517	1.852498
dataset=stoch	92	45	49.3	0.37909	0.856983
dataset=taylor	140	36	36.1	0.00022	0.000354

Chisq= 2.1 on 2 degrees of freedom, p= 0.3

However, if we construct a Generalized Wilcoxon Test (applies greater weight to early failure times), then we obtain R output 2. It shows p= 0.05 which can be considered marginally significant.

R output 2:

survdiff(formula = Surv(Time, Event) ~ dataset, data = datacamtaylorstock,
rho = 1)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
dataset=Cam	111	16.5	13.5	0.679	1.02
dataset=stoch	92	26.4	35.4	2.270	6.00
dataset=taylor	140	32.8	26.9	1.310	2.51

Chisq= 6 on 2 degrees of freedom, p= 0.05

3.2 Stratified Logrank Test

We might consider a cofounder exists. After investigation, we find that “iCluster” is a cofounder. Firstly, from R output 3 “iCluster” is noticed to be related to the survival. Secondly, it’s associated with “dataset” under the Pearson's Chi-squared test from R output 4.

R output 3:

survdiff(formula = Surv(Time, Event) ~ iCluster, data = datacamtaylorstock,
rho = 0)

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
iCluster=1	59	28	19.68	3.517	4.531
iCluster=2	81	11	26.33	8.924	12.440
iCluster=3	34	14	7.93	4.640	5.150
iCluster=4	49	10	13.42	0.872	1.038
iCluster=5	63	21	18.43	0.358	0.447
iCluster=6	21	8	2.70	10.377	10.946
iCluster=flat	22	3	6.50	1.887	2.093

Chisq= 31.3 on 6 degrees of freedom, p= 2e-05

R output 4:

Pearson's Chi-squared test
data: table(datacamtaylorstock$dataset, datacamtaylorstock$iCluster)
X-squared = 102.16, df = 12, p-value < 2.2e-16

Therefore, "iCluster" is considered as a cofounder and we will construct a Stratified Logrank Test among three datasets. Form R output 5, we can see p-value drops to 0.02 and it's undoubtedly significant under 0.05 level. So we consider the survival rates among these three datasets are significantly different after being stratified by "iCluster".

R output 5:

survdiff(formula = Surv(Time, Event) ~ dataset + strata(iCluster),
data = datacamtaylorstock, rho = 0)

n=329, 14 observations deleted due to missingness.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
dataset=Cam	104	18	13.1	1.840	2.83
dataset=stoch	85	41	51.1	2.010	7.73
dataset=taylor	140	36	30.8	0.889	2.21

Chisq= 8 on 2 degrees of freedom, p= 0.02

Section 4 Cox's proportional hazards regression model

4.1 Imputation of NA's

If we want to fit a Cox's model with the pooled datasets, we need to deal with missing values in these three datasets first. Since "Age", "PSA" and "ECE" are unavailable in all three datasets, we delete them. Besides, there are still 2 missing values in "Stage", 14 Missing values in "iCluster" and 4 missing values in "Gleason". We will impute all the NA's in "iCluster" and "Gleason" using Random Forest algorithm and delete the 2 NA's in "Stage".

4.2 Model 1: AIC Stepwise Cox's model using the pooled dataset with imputation

In this model, we consider all the variables and genes in the pooled dataset and use AIC stepwise technique to select significant variables. And the final model is

Surv(Time, Event, type = "right") ~ iCluster + Gleason + ENO2 + SYP + CHGA + SLC2A1 + SLC2A6 + SLC2A12 + SLC2A13 + SLC2A14 + HK1

From R output 6, we also notice most variables left in the model are significant.

R output 6:

	coef	*exp(coef)*	*se(coef)*	*z*	*p*
iCluster2	*-1.02e+00*	*3.62e-01*	*3.67e-01*	*-2.77*	*0.00565*
iCluster3	*-1.50e-01*	*8.61e-01*	*3.87e-01*	*-0.39*	*0.69796*
iCluster4	*-7.87e-01*	*4.55e-01*	*4.33e-01*	*-1.82*	*0.06908*
iCluster5	*-1.38e-01*	*8.71e-01*	*3.02e-01*	*-0.46*	*0.64688*
iCluster6	*3.62e-01*	*1.44e+00*	*4.82e-01*	*0.75*	*0.45299*
iClusterflat	*-5.28e-01*	*5.90e-01*	*6.96e-01*	*-0.76*	*0.44833*
Gleason6	*-3.99e+00*	*1.84e-02*	*1.21e+00*	*-3.30*	*0.00098*
Gleason7	*-2.12e+00*	*1.19e-01*	*1.09e+00*	*-1.95*	*0.05147*
Gleason8	*-1.22e+00*	*2.96e-01*	*1.12e+00*	*-1.08*	*0.27858*
Gleason9	*-1.80e+00*	*1.65e-01*	*1.10e+00*	*-1.64*	*0.10198*
Gleason5	*-1.82e+01*	*1.27e-08*	*2.36e+03*	*-0.01*	*0.99385*
ENO2	*9.98e-01*	*2.71e+00*	*4.57e-01*	*2.18*	*0.02906*
SYP	*7.47e-01*	*2.11e+00*	*2.93e-01*	*2.55*	*0.01073*
CHGA	*-2.33e-01*	*7.92e-01*	*1.56e-01*	*-1.50*	*0.13404*
SLC2A1	*1.50e+00*	*4.49e+00*	*5.05e-01*	*2.98*	*0.00292*

SLC2A6	*-1.18e+00*	*3.09e-01*	*6.33e-01*	*-1.86*	*0.06318*
SLC2A12	*-3.74e-01*	*6.88e-01*	*2.23e-01*	*-1.68*	*0.09317*
SLC2A13	*1.23e+00*	*3.42e+00*	*6.93e-01*	*1.77*	*0.07598*
SLC2A14	*-1.26e+00*	*2.85e-01*	*2.91e-01*	*-4.32*	*1.6e-05*
HK1	*-8.49e-01*	*4.28e-01*	*2.99e-01*	*-2.84*	*0.00454*

Likelihood ratio test=114.8 on 20 df, p=3e-15
n= 341, number of events= 98

4.2 Model 2: adding interaction terms between "dataset" and each gene

With the interaction terms between "dataset" and every gene, we obtain the final model after AIC stepwise selection as follows

Surv(Time, Event, type = "right") ~
iCluster + Gleason + dataset + ENO2 + NCAM1 + SYP + CHGA + FOLH1 + SLC2A1 + SLC2A3 + SLC2A4 + SLC2A5 + SLC2A6 + SLC2A7 + SLC2A8 + SLC2A10 + SLC2A12 + GCK + HK1 + dataset:NCAM1 + dataset:SYP + dataset:CHGA + dataset:FOLH1 + dataset:SLC2A1 + dataset:SLC2A3 + dataset:SLC2A5 + dataset:SLC2A6 + dataset:SLC2A7 + dataset:SLC2A8 + dataset:SLC2A12 + dataset:GCK + dataset:HK1

We can see interaction terms stay in the model. Likelihood ratio test=182.5 on 55 df and p=1e-15 from R output 7.

R output 7(omitting most parts due to the length):
Likelihood ratio test=182.5 on 55 df, p=1e-15
n= 341, number of events= 98

4.3 Model 3: comparison of Cox's models with and without imputation

Now, we fit a Cox's model without imputation(simply deleting 20 NA's) using the pooled dataset. We obtain the final model is

Surv(Time, Event, type = "right") ~
iCluster + Gleason + dataset + ENO2 + NCAM1 + SYP + CHGA + FOLH1 + SLC2A1 + SLC2A3 + SLC2A5 + SLC2A6 + SLC2A8 + SLC2A12 + GCK + HK1 + dataset:NCAM1 + dataset:SYP + dataset:CHGA + dataset:FOLH1 + dataset:SLC2A1 + dataset:SLC2A3 + dataset:SLC2A5 + dataset:SLC2A8 + dataset:GCK + dataset:HK1

We can see 20 interaction terms stay in the model. Likelihood ratio test=155.2 on 46 df and p=9e-14.

Compared this model with Model 2, Model 2 with imputation has larger likelihood ratio. Overall, the model with imputation(Model 2) seems better.

R output 8(omitting most parts due to the length):
Likelihood ratio test=155.2 on 46 df, p=9e-14
n= 323, number of events= 92

4.4 Model 4: Cox's model only using Cam dataset

Cam dataset is the only dataset containing all the variables including "Age", "PSA" and "ECE". Also, we know factors such as age are primary risk factors for prostate cancer survival. So we fit a Cox's model only using Cam dataset and obtain the final model is

Surv(Time, Event, type = "right") ~
Age + Stage + iCluster + Gleason + ECE + ENO2 + NCAM1 + SYP + FOLH1 + SLC2A1 + SLC2A3 + SLC2A4 + SLC2A5 + SLC2A7 + SLC2A8 + SLC2A10 + SLC2A12 + SLC2A14 + GCK + HK1 + HK2 + HK3

However, none of the variables are significant in the model even though the model is significant.

R output 9(omitting most parts due to the length):
Likelihood ratio test=121.8 on 33 df, p=4e-12
n= 103, number of events= 18

Section 5 Check proportionality assumption

We decide Model 2(with imputation and interaction terms using the pooled dataset) to be the best model. Now we check the proportionality assumption. From R output 10, we notice one level of “iCluster” and “dataset” are significant. So we may consider fitting the Cox’s model with strata of “iCluster” or “dataset”.

R output 10(omitting most parts due to the length):

	rho	chisq	p
iCluster3	0.240984	6.87e+00	0.00876
datasettaylor	0.217378	7.31e+00	0.00684

Section 6 Interpretation for Model 2

We need to include them when we interpret the model. For example, if a patient having the same demographic factors and genes comes from Stoch dataset instead of Cam dataset, then the hazard rate becomes $\exp(\beta_{datasetstoch} + \beta_{datasetstoch*NCAM1} * NCAM1 + \ldots + \beta_{datasetstoch*HK1} * HK1)$ as much.

Section 7 Investigating which factors and genes have greater effects using Random Survival Forests

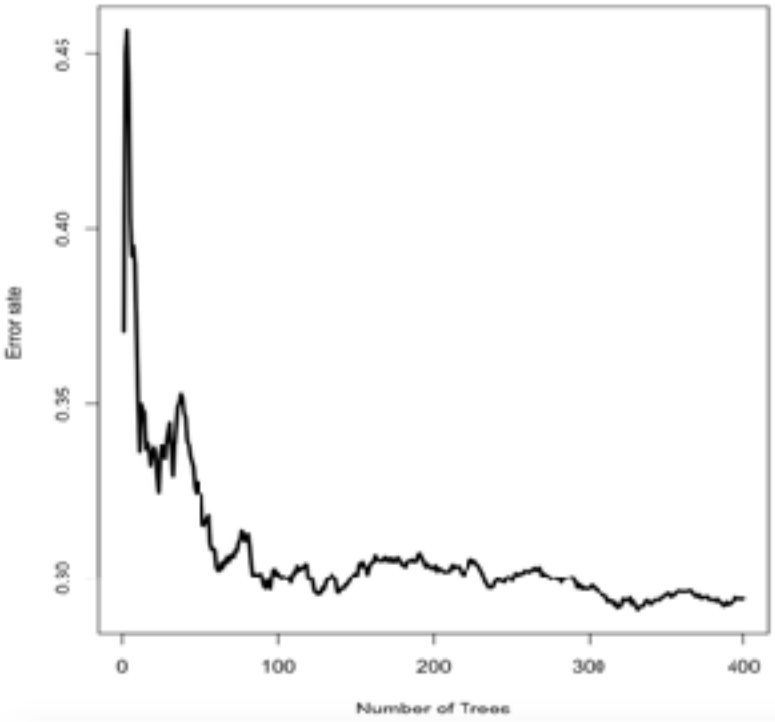
We can also use Random Survival Forest technique to find which factors and genes have greater effects on survival since it applies to right-censored data.

We grow 400 bootstrapped trees and implement log-rank splitting. It’s clear from Graph 3 that the Leave-one-out cross validation error drops fast and then stables at around 29% as the number of trees increases. And based on R output 11

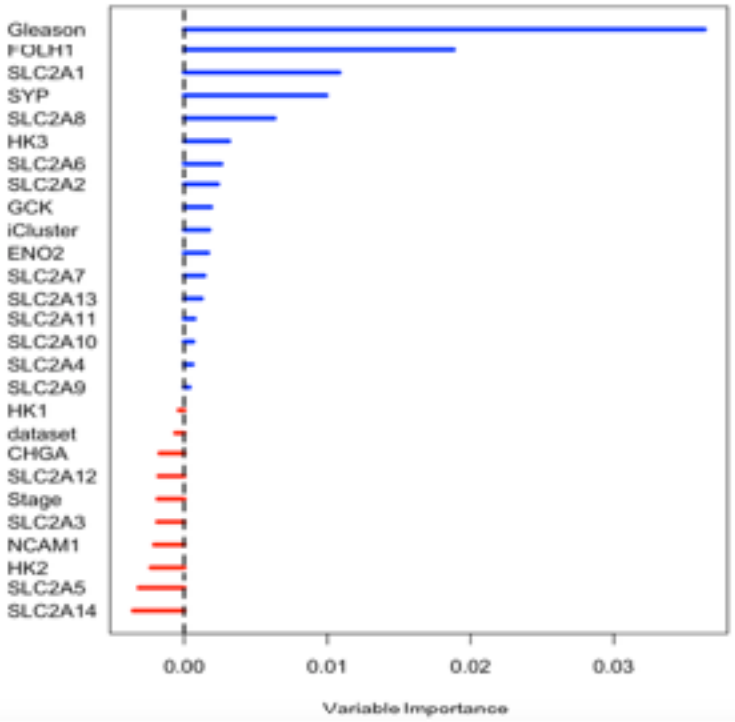
and Graph 4, we find the importance of each variable with ranking. “Gleason” has the greatest effect on survival followed by genes—FOLH1, SLC2A1, SYP, SLC2A8 and so forth.

R output 11:

	Importance	*Relative Importance*
Gleason	*0.0364*	*1.0000*
FOLH1	*0.0188*	*0.5171*
SLC2A1	*0.0108*	*0.2966*
SYP	*0.0099*	*0.2723*
SLC2A8	*0.0064*	*0.1746*
HK3	*0.0031*	*0.0839*
SLC2A6	*0.0025*	*0.0691*
SLC2A2	*0.0023*	*0.0627*
GCK	*0.0018*	*0.0499*
iCluster	*0.0017*	*0.0465*
ENO2	*0.0016*	*0.0444*
SLC2A7	*0.0014*	*0.0375*
SLC2A13	*0.0012*	*0.0324*
SLC2A11	*0.0007*	*0.0188*
SLC2A10	*0.0006*	*0.0164*
SLC2A4	*0.0005*	*0.0151*
SLC2A9	*0.0003*	*0.0095*
HK1	*-0.0004*	*-0.0096*
dataset	*-0.0006*	*-0.0163*
CHGA	*-0.0017*	*-0.0458*
SLC2A12	*-0.0017*	*-0.0476*
Stage	*-0.0018*	*-0.0495*
SLC2A3	*-0.0018*	*-0.0501*
NCAM1	*-0.0020*	*-0.0561*
HK2	*-0.0023*	*-0.0624*
SLC2A5	*-0.0031*	*-0.0853*
SLC2A14	*-0.0037*	*-0.1010*



Graph 3



Graph 4

Section 8 Comparing the results with Shiny App

In Table 2, we list the top 10 genes ranking from having the greatest to the least effect on time to BCR according to either their p-values or the importance. We can't directly and horizontally compare p-values or importance among these 5 methods since they use neither the same dataset nor the same techniques. Yet, we can vertically conclude within each method that

- Within Cam dataset in Shiny App, only the top 1 gene—SYP (p = 0.0011) has significant effect on BCR free time under level 0.05.
- Within Stoch dataset in Shiny App, the top four genes—SLC2A6, NCAM1, SLC2A9, SLC2A3 have significant effects on BCR free time under level 0.05.
- Within Taylor dataset in Shiny App, all the 10 genes— have significant effects on BCR free time under level 0.05.
- Within the pooled dataset using Random Survival Forest, the importance for the top 1 gene—FOLH1 is 1.88% followed by SLC2A1(1.08%), SYP(0.99%) and so forth.
- Within the pooled dataset using Cox's model(Model 2), the top four genes—ENO2, SYP, SLC2A4, SLC2A8 have significant effects on BCR free time under level 0.05.
- Overall, SYP, SLC2A1 and GCK show up in the top 10 list of every method.

Conclusion

Based on the data regarding the prostate cancer patients from three different datasets, we obtain the probabilities of freedom from BCR among three datasets are significantly different after stratification utilizing Stratified Logrank Test. Also, in order to investigate if demographic factors and genes have effects on survival, we build a Cox's proportional hazards regression model with imputation and interaction terms as well as a Random Survival Forest model. Furthermore, we compare the results with the previous research in a shiny app and three genes(SYP, SLC2A1 and GCK) appear to have effects on Time to BCR in all 5 methods.

Table 2. Comparison of ranking of genes having effects on survival

Ranking	From Cam dataset in Shiny App		From Stoch dataset in Shiny App		From Taylor dataset in Shiny App		Form pooled dataset with Random Survival Forest		Form pooled dataset with Cox's model (Model 2)	
	Gene	p.value	Gene	p.value	Gene	p.value	Gene	Importance	Gene	p.value
1	SYP	0.0011	SLC2A6	0.0072	HK1	0.0021	FOLH1	0.0188	ENO2	0.00291
2	SLC2A1	0.14	NCAM1	0.0098	SLC2A1	0.0026	SLC2A1	0.0108	SYP	0.00305
3	GCK	0.21	SLC2A9	0.042	SLC2A3	0.0042	SYP	0.0099	SLC2A4	0.02993
4	SLC2A13	0.25	SLC2A3	0.048	SLC2A14	0.0056	SLC2A8	0.0064	SLC2A8	0.03114
5	ENO2	0.26	SLC2A8	0.11	SYP	0.0075	HK3	0.0031	SLC2A10	0.06023
6	CHGA	0.28	SYP	0.14	SLC2A12	0.012	SLC2A6	0.0025	NCAM1	0.07259
7	HK1	0.29	GCK	0.14	GCK	0.015	SLC2A2	0.0023	HK1	0.08110
8	FOLH1	0.31	SLC2A14	0.15	SLC2A2	0.019	GCK	0.0018	GCK	0.11131
9	SLC2A8	0.33	SLC2A1	0.18	CHGA	0.027	ENO2	0.0016	SLC2A1	0.15673
10	SLC2A4	0.36	HK2	0.24	HK3	0.036	SLC2A7	0.0014	CHGA	0.17979

Reference

1. Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, et al. (May 1997). "The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews". The New England Journal of Medicine. 336 (20): 1401–8.

2. Gallagher RP, Fleshner N (October 1998). "Prostate cancer: 3. Individual risk factors" (PDF). CMAJ. 159 (7): 807–13. PMC 1232741. PMID 9805030. Archived (PDF) from the original on 2009-12-29.

3. Beuzeboc P, Soulié M, Richaud P, Salomon L, Staerman F, Peyromaure M, et al. (December 2009). "[Fusion genes and prostate cancer. From discovery to prognosis and therapeutic perspectives]". Progres en Urologie (in French). 19 (11): 819–24.

4. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. (March 2008). "Multiple newly identified loci associated with prostate cancer susceptibility". Nature Genetics. 40 (3): 316–21.

5. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, et al. (March 2008). "Multiple loci identified in a genome-wide association study of prostate cancer". Nature Genetics. 40 (3): 310–5.