OCM_modelbuilding

Elizabeth Chase

7/20/2020

Descriptives

First, an unweighted table of descriptives:

```
table_unweight <-
  CreateTableOne(
    vars = c(
      "age",
      "SDDSRVYR",
      "race2",
      "military",
      "usbirth",
      "educ",
      "marital2",
      "INDFMPIR",
      "insurance",
      "care",
      "hospital",
      "hypertension",
      "highchol",
      "diabetic",
      "mental",
      "kidney",
      "asthma",
      "anemic",
      "arthritis",
      "chf",
      "chd",
      "angina",
      "mi",
      "mi_chd",
      "stroke",
      "emphysema",
      "bronch",
      "liver",
      "hiv",
      "pc",
      "agepc",
      "BMXBMI",
      "underweight",
      "overweight_ex",
      "overweight2",
```

```
      "obese",
      "obese2",
      "alcoholic",
      "smoker",
      "permth_exm"
    ),
    strata = "mortstat",
    includeNA = TRUE,
    data = mydata[mydata$inmodel1 == 1, ],
    argsNormal = list(NULL),
    argsNonNormal = list(var.equal = TRUE)
  )

kable(print(
  table_unweight,
  contDigits = 1,
  varLabel = T,
  printToggle = FALSE
),
booktabs = T) %>% add_indent(c(4:9, 11:13, 16:19, 21:25, 27:29, 52:54, 65:67))
```

	0	1	p	test
n	1961	459		
Age (years) (mean (SD))	56.8 (10.8)	70.3 (11.4)	<0.001	
Survey Year (%)			NaN	
1999-2000	0 (0.0)	0 (0.0)		
2001-2002	308 (15.7)	114 (24.8)		
2003-2004	261 (13.3)	126 (27.5)		
2005-2006	353 (18.0)	78 (17.0)		
2007-2008	476 (24.3)	88 (19.2)		
2009-2010	563 (28.7)	53 (11.5)		
Race (%)			<0.001	
NHB	343 (17.5)	68 (14.8)		
NHW	1058 (54.0)	304 (66.2)		
Other	560 (28.6)	87 (19.0)		
Military service = Did not serve (%)	1296 (66.1)	190 (41.4)	<0.001	
Birthplace (%)			<0.001	
Born U.S.	732 (37.3)	280 (61.0)		
Born Mexico	109 (5.6)	22 (4.8)		
Born elsewhere	81 (4.1)	16 (3.5)		
NA	1039 (53.0)	141 (30.7)		
Education (%)			<0.001	
Less than 9th grade	222 (11.3)	83 (18.1)		
9th-11th grade	226 (11.5)	69 (15.0)		
HS graduate	476 (24.3)	115 (25.1)		
Some college	500 (25.5)	107 (23.3)		
College graduate	537 (27.4)	85 (18.5)		
Marital status (%)			<0.001	
Married	1515 (77.3)	298 (64.9)		
Separated	323 (16.5)	143 (31.2)		
Single	123 (6.3)	18 (3.9)		
Income:Poverty Level Ratio (mean (SD))	3.2 (1.6)	2.5 (1.5)	<0.001	
Has insurance = Yes (%)	1657 (84.5)	423 (92.2)	<0.001	
Has primary care provider = Yes (%)	1731 (88.3)	432 (94.1)	<0.001	
Hospitalized in last 12 months = Yes (%)	184 (9.4)	95 (20.7)	<0.001	
Has hypertension = Yes (%)	893 (45.5)	276 (60.1)	<0.001	
Has high cholesterol = Yes (%)	1135 (57.9)	232 (50.5)	0.005	
Diabetic = Yes (%)	346 (17.6)	107 (23.3)	0.006	
Has kidney disease = Yes (%)	30 (1.5)	25 (5.4)	<0.001	
Has asthma = Yes (%)	194 (9.9)	42 (9.2)	0.693	
Anemic = Yes (%)	29 (1.5)	22 (4.8)	<0.001	
Arthritic = Yes (%)	612 (31.2)	201 (43.8)	<0.001	
Has congestive heart failure = Yes (%)	54 (2.8)	58 (12.6)	<0.001	
Has coronary heart disease = Yes (%)	129 (6.6)	87 (19.0)	<0.001	
Has angina = Yes (%)	58 (3.0)	55 (12.0)	<0.001	
Previous heart attack = Yes (%)	104 (5.3)	79 (17.2)	<0.001	
Previous heart attack or CHD = Yes (%)	166 (8.5)	115 (25.1)	<0.001	
Previous stroke = Yes (%)	50 (2.5)	45 (9.8)	<0.001	
Has emphysema = Yes (%)	35 (1.8)	40 (8.7)	<0.001	
Has chronic bronchitis = Yes (%)	80 (4.1)	31 (6.8)	0.019	
Has liver disease = Yes (%)	92 (4.7)	21 (4.6)	1.000	
Has HIV (%)			<0.001	
No	753 (38.4)	34 (7.4)		
Yes	5 (0.3)	1 (0.2)		
NA	1203 (61.3)	424 (92.4)		
Has prostate cancer = PC (%)	74 (3.8)	53 (11.5)	<0.001	
Age of PC Diagnosis (mean (SD))	66.3 (7.2)	71.0 (7.9)	0.001	

And now the same table, but survey weighted:

```
table_weight <-
  svyCreateTableOne(
    vars = c(
      "age",
      "SDDSRVYR",
      "race2",
      "military",
      "usbirth",
      "educ",
      "marital2",
      "INDFMPIR",
      "insurance",
      "care",
      "hospital",
      "hypertension",
      "highchol",
      "diabetic",
      "mental",
      "kidney",
      "asthma",
      "anemic",
      "arthritis",
      "chf",
      "chd",
      "angina",
      "mi",
      "mi_chd",
      "stroke",
      "emphysema",
      "bronch",
      "liver",
      "hiv",
      "pc",
      "agepc",
      "BMXBMI",
      "underweight",
      "overweight_ex",
      "overweight2",
      "obese",
      "obese2",
      "alcoholic",
      "smoker",
      "permth_exm"
    ),
    strata = "mortstat",
    includeNA = TRUE,
    data = nhanes_sub,
    argsNormal = list(NULL),
    argsNonNormal = list(var.equal = TRUE)
  )

kable(print(
```

```
  table_weight,
  contDigits = 1,
  varLabel = T,
  printToggle = FALSE
),
booktabs = T) %>% add_indent(c(4:9, 11:13, 16:19, 21:25, 27:29, 52:54, 65:67))
```

	0	1	p	test
n	83364634.6	11793391.1		
age (mean (SD))	53.9 (9.6)	67.2 (12.0)	<0.001	
SDDSRVYR (%)			NaN	
1999-2000	0.0 (0.0)	0.0 (0.0)		
2001-2002	13637483.5 (16.4)	2817758.2 (23.9)		
2003-2004	12498153.0 (15.0)	3443393.9 (29.2)		
2005-2006	18298272.9 (21.9)	2172599.8 (18.4)		
2007-2008	18466663.7 (22.2)	1952828.3 (16.6)		
2009-2010	20464061.5 (24.5)	1406810.8 (11.9)		
race2 (%)			0.179	
NHB	6616210.2 (7.9)	1000001.1 (8.5)		
NHW	65688832.0 (78.8)	9619031.9 (81.6)		
Other	11059592.5 (13.3)	1174358.0 (10.0)		
military = Did not serve (%)	56759467.5 (68.1)	4916180.4 (41.7)	<0.001	
usbirth (%)			NA	
Born U.S.	39095146.6 (46.9)	7964869.4 (67.5)		
Born Mexico	1495715.9 (1.8)	125140.7 (1.1)		
Born elsewhere	3843046.9 (4.6)	343741.9 (2.9)		
NA	38930725.2 (46.7)	3359639.2 (28.5)		
educ (%)			<0.001	
Less than 9th grade	3708271.5 (4.4)	1333108.3 (11.3)		
9th-11th grade	6453932.2 (7.7)	1665286.0 (14.1)		
HS graduate	20551630.1 (24.7)	3056911.8 (25.9)		
Some college	24294518.3 (29.1)	3450405.3 (29.3)		
College graduate	28356282.5 (34.0)	2287679.8 (19.4)		
marital2 (%)			<0.001	
Married	66023841.8 (79.2)	7708866.3 (65.4)		
Separated	12199918.9 (14.6)	3571060.8 (30.3)		
Single	5140873.9 (6.2)	513464.0 (4.4)		
INDFMPIR (mean (SD))	3.7 (1.5)	2.8 (1.5)	<0.001	
insurance = Yes (%)	74437872.3 (89.3)	10852487.5 (92.0)	0.188	
care = Yes (%)	74621492.2 (89.5)	10998922.1 (93.3)	0.089	
hospital = Yes (%)	6613356.7 (7.9)	2148591.9 (18.2)	<0.001	
hypertension = Yes (%)	34557037.5 (41.5)	6948686.9 (58.9)	<0.001	
highchol = Yes (%)	47082705.5 (56.5)	6084061.3 (51.6)	0.154	
diabetic = Yes (%)	10444444.4 (12.5)	2568461.5 (21.8)	<0.001	
kidney = Yes (%)	931459.8 (1.1)	497892.8 (4.2)	<0.001	
asthma = Yes (%)	8656223.8 (10.4)	1212122.2 (10.3)	0.949	
anemic = Yes (%)	841606.9 (1.0)	436580.8 (3.7)	<0.001	
arthritis = Yes (%)	23265414.6 (27.9)	5325542.8 (45.2)	<0.001	
chf = Yes (%)	1454915.9 (1.7)	1345319.9 (11.4)	<0.001	
chd = Yes (%)	4589740.9 (5.5)	2180089.5 (18.5)	<0.001	
angina = Yes (%)	2579917.9 (3.1)	1207491.9 (10.2)	<0.001	
mi = Yes (%)	3726339.3 (4.5)	1954138.4 (16.6)	<0.001	
mi_chd = Yes (%)	5913621.8 (7.1)	2794710.5 (23.7)	<0.001	
stroke = Yes (%)	1327518.8 (1.6)	1115904.2 (9.5)	<0.001	
emphysema = Yes (%)	1034810.7 (1.2)	992237.3 (8.4)	<0.001	
bronch = Yes (%)	3244248.0 (3.9)	721125.7 (6.1)	0.091	
liver = Yes (%)	3909809.6 (4.7)	479437.9 (4.1)	0.561	
hiv (%)			NA	
No	39133626.8 (46.9)	1308221.2 (11.1)		
Yes	252279.6 (0.3)	33121.3 (0.3)		
NA	43978728.2 (52.8)	10452048.6 (88.6)		
pc = PC (%)	1950111.0 (2.3)	1290483.3 (10.9)	<0.001	
cvno (mean (SD))	65.2 (7.7)	69.8 (7.8)	0.024	

Cox Model Building, With Survey Weights

We start by fitting a Cox model with all of our predictors, using the NHANES survey weights.

```
nhanes_cox_full <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + alcoholic + anemic +
      arthritis + asthma + bronch + diabetic + educ + emphysema + hypertension +
      marital2 + mi + liver + race2 + smoker + stroke + underweight + overweight_ex +
      obese + angina + chf + chd + care + hospital + highchol + insurance + kidney +
      military + pc,
    design = nhanes_sub
  )

summary(nhanes_cox_full)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (153) clusters.
## subset(nhanes_svy, inmodel1 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     alcoholic + anemic + arthritis + asthma + bronch + diabetic +
##     educ + emphysema + hypertension + marital2 + mi + liver +
##     race2 + smoker + stroke + underweight + overweight_ex + obese +
##     angina + chf + chd + care + hospital + highchol + insurance +
##     kidney + military + pc, design = nhanes_sub)
##
##   n= 2420, number of events= 459
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## scale(age, scale = FALSE)  0.096155  1.100929  0.008075 11.907  < 2e-16 ***
## alcoholicYes               0.072048  1.074707  0.155844  0.462 0.643859
## anemicYes                 -0.183003  0.832766  0.238155 -0.768 0.442238
## arthritisYes               0.054956  1.056494  0.135727  0.405 0.685551
## asthmaYes                  0.075647  1.078582  0.193412  0.391 0.695710
## bronchYes                 -0.023778  0.976503  0.265525 -0.090 0.928645
## diabeticYes                0.416204  1.516195  0.132124  3.150 0.001632 **
## educ9th-11th grade         0.031828  1.032340  0.211974  0.150 0.880645
## educHS graduate           -0.263940  0.768019  0.198890 -1.327 0.184486
## educSome college          -0.163862  0.848859  0.184065 -0.890 0.373337
## educCollege graduate      -0.688882  0.502137  0.213738 -3.223 0.001268 **
## emphysemaYes               0.364086  1.439197  0.202647  1.797 0.072392 .
## hypertensionYes            0.127122  1.135555  0.122298  1.039 0.298598
## marital2Separated          0.534344  1.706328  0.107188  4.985 6.19e-07 ***
## marital2Single             0.435140  1.545179  0.276599  1.573 0.115677
## miYes                      0.125422  1.133627  0.158919  0.789 0.429983
## liverYes                  -0.076689  0.926178  0.221500 -0.346 0.729172
## race2NHW                   0.128354  1.136955  0.193965  0.662 0.508140
## race2Other                -0.046508  0.954557  0.267190 -0.174 0.861817
## smokerCurrent              0.977610  2.658096  0.166306  5.878 4.14e-09 ***
## smokerFormer               0.527470  1.694640  0.134821  3.912 9.14e-05 ***
## strokeYes                  0.686014  1.985784  0.193506  3.545 0.000392 ***
## underweightYes             1.494623  4.457654  0.487766  3.064 0.002182 **
## overweight_exYes           0.080505  1.083834  0.166965  0.482 0.629687
## obeseYes                   0.062798  1.064811  0.178104  0.353 0.724396
```

```
## anginaYes                   0.150026  1.161865  0.174782  0.858 0.390693
## chfYes                      0.778887  2.179045  0.163455  4.765 1.89e-06 ***
## chdYes                     -0.072033  0.930500  0.163797 -0.440 0.660102
## careYes                    -0.354338  0.701638  0.301839 -1.174 0.240424
## hospitalYes                 0.150703  1.162652  0.170904  0.882 0.377883
## highcholYes                -0.234186  0.791215  0.127234 -1.841 0.065680 .
## insuranceYes               -0.112722  0.893399  0.271715 -0.415 0.678248
## kidneyYes                   0.570171  1.768569  0.366845  1.554 0.120124
## militaryDid not serve       0.098652  1.103683  0.146811  0.672 0.501604
## pcPC                        0.253581  1.288632  0.188437  1.346 0.178397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## scale(age, scale = FALSE)    1.1009     0.9083    1.0836    1.1185
## alcoholicYes                 1.0747     0.9305    0.7918    1.4586
## anemicYes                    0.8328     1.2008    0.5222    1.3281
## arthritisYes                 1.0565     0.9465    0.8097    1.3785
## asthmaYes                    1.0786     0.9271    0.7383    1.5757
## bronchYes                    0.9765     1.0241    0.5803    1.6432
## diabeticYes                  1.5162     0.6595    1.1703    1.9644
## educ9th-11th grade           1.0323     0.9687    0.6814    1.5641
## educHS graduate              0.7680     1.3021    0.5201    1.1341
## educSome college             0.8489     1.1781    0.5918    1.2176
## educCollege graduate         0.5021     1.9915    0.3303    0.7634
## emphysemaYes                 1.4392     0.6948    0.9674    2.1410
## hypertensionYes              1.1356     0.8806    0.8935    1.4431
## marital2Separated            1.7063     0.5861    1.3830    2.1052
## marital2Single               1.5452     0.6472    0.8985    2.6572
## miYes                        1.1336     0.8821    0.8302    1.5479
## liverYes                     0.9262     1.0797    0.6000    1.4297
## race2NHW                     1.1370     0.8795    0.7774    1.6628
## race2Other                   0.9546     1.0476    0.5654    1.6115
## smokerCurrent                2.6581     0.3762    1.9187    3.6824
## smokerFormer                 1.6946     0.5901    1.3011    2.2072
## strokeYes                    1.9858     0.5036    1.3590    2.9016
## underweightYes               4.4577     0.2243    1.7136   11.5956
## overweight_exYes             1.0838     0.9227    0.7813    1.5034
## obeseYes                     1.0648     0.9391    0.7511    1.5096
## anginaYes                    1.1619     0.8607    0.8249    1.6366
## chfYes                       2.1790     0.4589    1.5817    3.0019
## chdYes                       0.9305     1.0747    0.6750    1.2827
## careYes                      0.7016     1.4252    0.3883    1.2678
## hospitalYes                  1.1627     0.8601    0.8317    1.6253
## highcholYes                  0.7912     1.2639    0.6166    1.0153
## insuranceYes                 0.8934     1.1193    0.5245    1.5217
## kidneyYes                    1.7686     0.5654    0.8617    3.6298
## militaryDid not serve        1.1037     0.9061    0.8277    1.4717
## pcPC                         1.2886     0.7760    0.8907    1.8643
##
## Concordance= 0.842  (se = 0.014 )
## Likelihood ratio test= NA  on 35 df,   p=NA
## Wald test            = 1487  on 35 df,   p=<2e-16
```

```
## Score (logrank) test = NA  on 35 df,   p=NA
```

And now we fit the same model, but using only those predictors that also appear in PLCO, our validation data:

```
plco_cox <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + arthritis + bronch +
      diabetic + educ + emphysema + hypertension + marital2 + mi_chd + underweight +
      overweight_ex + obese + liver + race2 + smoker + stroke,
    design = nhanes_sub
  )

summary(plco_cox)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (153) clusters.
## subset(nhanes_svy, inmodel1 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     arthritis + bronch + diabetic + educ + emphysema + hypertension +
##     marital2 + mi_chd + underweight + overweight_ex + obese +
##     liver + race2 + smoker + stroke, design = nhanes_sub)
##
##   n= 2420, number of events= 459
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## scale(age, scale = FALSE)  0.093525  1.098038  0.006777 13.799  < 2e-16 ***
## arthritisYes               0.046423  1.047517  0.133934  0.347 0.728884
## bronchYes                  0.091737  1.096076  0.261648  0.351 0.725881
## diabeticYes                0.446821  1.563335  0.136367  3.277 0.001051 **
## educ9th-11th grade        -0.047528  0.953583  0.220745 -0.215 0.829526
## educHS graduate           -0.346511  0.707151  0.192520 -1.800 0.071881 .
## educSome college          -0.239592  0.786949  0.182707 -1.311 0.189741
## educCollege graduate      -0.780836  0.458023  0.221052 -3.532 0.000412 ***
## emphysemaYes               0.352222  1.422225  0.220519  1.597 0.110211
## hypertensionYes            0.134036  1.143435  0.121447  1.104 0.269739
## marital2Separated          0.612786  1.845567  0.097686  6.273 3.54e-10 ***
## marital2Single             0.485915  1.625661  0.285815  1.700 0.089112 .
## mi_chdYes                  0.263708  1.301748  0.112025  2.354 0.018572 *
## underweightYes             1.503103  4.495618  0.494850  3.037 0.002386 **
## overweight_exYes           0.108068  1.114123  0.170526  0.634 0.526257
## obeseYes                   0.091432  1.095742  0.186600  0.490 0.624143
## liverYes                   0.033962  1.034545  0.210024  0.162 0.871538
## race2NHW                   0.097423  1.102327  0.200636  0.486 0.627271
## race2Other                -0.106175  0.899267  0.271521 -0.391 0.695768
## smokerCurrent              1.011783  2.750502  0.178649  5.664 1.48e-08 ***
## smokerFormer               0.534151  1.706000  0.141049  3.787 0.000152 ***
## strokeYes                  0.753928  2.125332  0.179921  4.190 2.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                            exp(coef) exp(-coef) lower .95 upper .95
## scale(age, scale = FALSE)     1.0980     0.9107    1.0835    1.1127
## arthritisYes                  1.0475     0.9546    0.8057    1.3620
```

```
## bronchYes                    1.0961     0.9123    0.6563    1.8304
## diabeticYes                  1.5633     0.6397    1.1967    2.0423
## educ9th-11th grade           0.9536     1.0487    0.6187    1.4698
## educHS graduate              0.7072     1.4141    0.4849    1.0313
## educSome college             0.7869     1.2707    0.5501    1.1258
## educCollege graduate         0.4580     2.1833    0.2970    0.7064
## emphysemaYes                 1.4222     0.7031    0.9231    2.1912
## hypertensionYes              1.1434     0.8746    0.9012    1.4507
## marital2Separated            1.8456     0.5418    1.5240    2.2350
## marital2Single               1.6257     0.6151    0.9284    2.8465
## mi_chdYes                    1.3017     0.7682    1.0451    1.6214
## underweightYes               4.4956     0.2224    1.7044   11.8579
## overweight_exYes             1.1141     0.8976    0.7976    1.5563
## obeseYes                     1.0957     0.9126    0.7601    1.5796
## liverYes                     1.0345     0.9666    0.6855    1.5614
## race2NHW                     1.1023     0.9072    0.7439    1.6334
## race2Other                   0.8993     1.1120    0.5282    1.5311
## smokerCurrent                2.7505     0.3636    1.9380    3.9037
## smokerFormer                 1.7060     0.5862    1.2940    2.2493
## strokeYes                    2.1253     0.4705    1.4938    3.0240
##
## Concordance= 0.837  (se = 0.014 )
## Likelihood ratio test= NA  on 22 df,   p=NA
## Wald test            = 1008  on 22 df,   p=<2e-16
## Score (logrank) test = NA  on 22 df,   p=NA
```

From this, we see that model performance is not overly hurt by restricting to PLCO-NHANES common covariates (C-index of 0.84 vs. 0.83). We try dropping the covariates arthritis, bronchitis, liver disease, and race, because these are overwhelmingly nonsignificant. In addition, we try dropping emphysema, because this may be more challenging to obtain in a clinical setting:

```
plco_red1 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + mi_chd + underweight + overweight_ex + obese +
      smoker + stroke,
    design = nhanes_sub
  )

summary(plco_red1)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (153) clusters.
## subset(nhanes_svy, inmodel1 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + mi_chd + underweight +
##     overweight_ex + obese + smoker + stroke, design = nhanes_sub)
##
##   n= 2420, number of events= 459
##
##                               coef exp(coef)  se(coef)      z Pr(>|z|)
## scale(age, scale = FALSE)  0.095443  1.100146  0.006307 15.132  < 2e-16 ***
## diabeticYes                0.430218  1.537593  0.134715  3.194 0.001405 **
## educ9th-11th grade         0.017692  1.017850  0.212777  0.083 0.933733
```

```
## educHS graduate           -0.306187  0.736249  0.187165 -1.636 0.101857
## educSome college          -0.189945  0.827005  0.166556 -1.140 0.254109
## educCollege graduate      -0.728186  0.482784  0.212519 -3.426 0.000612 ***
## hypertensionYes            0.134736  1.144234  0.120191  1.121 0.262280
## marital2Separated          0.617589  1.854452  0.097926  6.307 2.85e-10 ***
## marital2Single             0.492308  1.636088  0.282256  1.744 0.081126 .
## mi_chdYes                  0.311169  1.365019  0.107157  2.904 0.003686 **
## underweightYes             1.496002  4.463807  0.486201  3.077 0.002092 **
## overweight_exYes           0.115129  1.122019  0.164597  0.699 0.484264
## obeseYes                   0.104885  1.110583  0.180088  0.582 0.560292
## smokerCurrent              1.040176  2.829714  0.173852  5.983 2.19e-09 ***
## smokerFormer               0.559612  1.749993  0.137544  4.069 4.73e-05 ***
## strokeYes                  0.753564  2.124559  0.176240  4.276 1.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## scale(age, scale = FALSE)    1.1001     0.9090    1.0866    1.1138
## diabeticYes                  1.5376     0.6504    1.1808    2.0022
## educ9th-11th grade           1.0178     0.9825    0.6708    1.5445
## educHS graduate              0.7362     1.3582    0.5102    1.0625
## educSome college             0.8270     1.2092    0.5967    1.1463
## educCollege graduate         0.4828     2.0713    0.3183    0.7322
## hypertensionYes              1.1442     0.8739    0.9041    1.4482
## marital2Separated            1.8545     0.5392    1.5306    2.2468
## marital2Single               1.6361     0.6112    0.9409    2.8449
## mi_chdYes                    1.3650     0.7326    1.1064    1.6840
## underweightYes               4.4638     0.2240    1.7213   11.5761
## overweight_exYes             1.1220     0.8913    0.8126    1.5492
## obeseYes                     1.1106     0.9004    0.7803    1.5807
## smokerCurrent                2.8297     0.3534    2.0126    3.9786
## smokerFormer                 1.7500     0.5714    1.3365    2.2915
## strokeYes                    2.1246     0.4707    1.5040    3.0011
##
## Concordance= 0.836  (se = 0.014 )
## Likelihood ratio test= NA  on 16 df,   p=NA
## Wald test            = 887.6  on 16 df,   p=<2e-16
## Score (logrank) test = NA  on 16 df,   p=NA
```

That had minimal effect on predictive performance, so let's stick with that. We now expand our sample out to only include individuals with complete data for our 9 current covariates (age, diabetes, education, hypertension, marital status, MI/CHD, BMI, smoking status, stroke), to increase our sample size. We consider interactions between all variables and age.

```
mydata$inmodel2 <-
  ifelse(
    mydata$sex == "Male" &
      mydata$age > 40 &
      (mydata$cancer == 0 |
         mydata$pc == "PC") &
      !is.na(mydata$age) &
      !is.na(mydata$diabetic) &
      !is.na(mydata$educ) &
      !is.na(mydata$hypertension) &
```

```
      !is.na(mydata$marital2) &
      !is.na(mydata$mi_chd) &
      !is.na(mydata$BMXBMI) &
      !is.na(mydata$smoker) &
      !is.na(mydata$stroke) &
      !is.na(mydata$mortstat) & !is.na(mydata$pc),
    1,
    0
  )

nhanes_svy <-
  svydesign(
    data = mydata,
    id =  ~ psu,
    strata =  ~ strata,
    weights =  ~ sweights,
    nest = TRUE
  )
nhanes_sub2 <- subset(nhanes_svy, inmodel2 == 1)

plco_moredat <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + mi_chd + underweight + overweight_ex + obese +
      smoker + stroke,
    design = nhanes_sub2
  )

summary(plco_moredat)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + mi_chd + underweight +
##     overweight_ex + obese + smoker + stroke, design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## scale(age, scale = FALSE)  0.090033  1.094210  0.003042 29.597  < 2e-16 ***
## diabeticYes                0.405676  1.500316  0.080302  5.052 4.37e-07 ***
## educ9th-11th grade         0.115611  1.122559  0.073464  1.574 0.115556
## educHS graduate           -0.118053  0.888649  0.095412 -1.237 0.215975
## educSome college          -0.164243  0.848536  0.095436 -1.721 0.085254 .
## educCollege graduate      -0.539707  0.582919  0.080002 -6.746 1.52e-11 ***
## hypertensionYes            0.225645  1.253131  0.064545  3.496 0.000472 ***
## marital2Separated          0.402142  1.495024  0.061403  6.549 5.78e-11 ***
## marital2Single             0.604205  1.829797  0.135474  4.460 8.20e-06 ***
## mi_chdYes                  0.295650  1.344000  0.065417  4.519 6.20e-06 ***
## underweightYes             0.894937  2.447181  0.284184  3.149 0.001638 **
## overweight_exYes          -0.123725  0.883622  0.073364 -1.686 0.091705 .
## obeseYes                   0.026280  1.026629  0.083395  0.315 0.752663
```

```
## smokerCurrent               0.823608  2.278706  0.092652  8.889  < 2e-16 ***
## smokerFormer                0.188388  1.207302  0.067052  2.810 0.004961 **
## strokeYes                   0.517356  1.677586  0.091884  5.631 1.80e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## scale(age, scale = FALSE)    1.0942     0.9139    1.0877    1.1008
## diabeticYes                  1.5003     0.6665    1.2818    1.7560
## educ9th-11th grade           1.1226     0.8908    0.9720    1.2964
## educHS graduate              0.8886     1.1253    0.7371    1.0714
## educSome college             0.8485     1.1785    0.7038    1.0231
## educCollege graduate         0.5829     1.7155    0.4983    0.6819
## hypertensionYes              1.2531     0.7980    1.1042    1.4221
## marital2Separated            1.4950     0.6689    1.3255    1.6862
## marital2Single               1.8298     0.5465    1.4031    2.3863
## mi_chdYes                    1.3440     0.7440    1.1823    1.5279
## underweightYes               2.4472     0.4086    1.4021    4.2713
## overweight_exYes             0.8836     1.1317    0.7653    1.0203
## obeseYes                     1.0266     0.9741    0.8718    1.2089
## smokerCurrent                2.2787     0.4388    1.9003    2.7325
## smokerFormer                 1.2073     0.8283    1.0586    1.3769
## strokeYes                    1.6776     0.5961    1.4011    2.0086
##
## Concordance= 0.817  (se = 0.008 )
## Likelihood ratio test= NA  on 16 df,   p=NA
## Wald test            = 2285  on 16 df,   p=<2e-16
## Score (logrank) test = NA  on 16 df,   p=NA

plco_int <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + mi_chd + underweight + overweight_ex + obese +
      smoker + stroke + scale(age, scale = FALSE) * diabetic +
      scale(age, scale = FALSE) * educ + scale(age, scale = FALSE) * hypertension +
      scale(age, scale = FALSE) * marital2 + scale(age, scale = FALSE) * mi_chd +
      scale(age, scale = FALSE) * underweight + scale(age, scale = FALSE) * overweight_ex +
      scale(age, scale = FALSE) * obese + scale(age, scale = FALSE) * smoker +
      scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + mi_chd + underweight +
##     overweight_ex + obese + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     marital2 + scale(age, scale = FALSE) * mi_chd + scale(age,
```

```
##     scale = FALSE) * underweight + scale(age, scale = FALSE) *
##     overweight_ex + scale(age, scale = FALSE) * obese + scale(age,
##     scale = FALSE) * smoker + scale(age, scale = FALSE) * stroke,
##     design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                                    coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                      0.117426  1.124598  0.010817
## diabeticYes                                    0.571250  1.770479  0.108024
## educ9th-11th grade                             0.004591  1.004602  0.114560
## educHS graduate                               -0.181569  0.833960  0.136524
## educSome college                              -0.210595  0.810102  0.128048
## educCollege graduate                          -0.654639  0.519629  0.148970
## hypertensionYes                                0.321716  1.379492  0.086673
## marital2Separated                              0.516196  1.675641  0.088398
## marital2Single                                 0.585157  1.795272  0.138442
## mi_chdYes                                      0.394706  1.483948  0.105914
## underweightYes                                 0.965249  2.625442  0.312938
## overweight_exYes                              -0.090350  0.913611  0.114034
## obeseYes                                       0.065868  1.068086  0.119467
## smokerCurrent                                  0.858627  2.359919  0.111931
## smokerFormer                                   0.213578  1.238100  0.118778
## strokeYes                                      0.630441  1.878438  0.158679
## scale(age, scale = FALSE):diabeticYes         -0.025179  0.975136  0.008057
## scale(age, scale = FALSE):educ9th-11th grade   0.009944  1.009993  0.008894
## scale(age, scale = FALSE):educHS graduate      0.004781  1.004792  0.008847
## scale(age, scale = FALSE):educSome college     0.001823  1.001824  0.008224
## scale(age, scale = FALSE):educCollege graduate 0.014642  1.014750  0.010694
## scale(age, scale = FALSE):hypertensionYes     -0.019485  0.980704  0.006103
## scale(age, scale = FALSE):marital2Separated   -0.012127  0.987947  0.006879
## scale(age, scale = FALSE):marital2Single      -0.010374  0.989680  0.011214
## scale(age, scale = FALSE):mi_chdYes           -0.010227  0.989826  0.007980
## scale(age, scale = FALSE):underweightYes      -0.017083  0.983062  0.028178
## scale(age, scale = FALSE):overweight_exYes    -0.001464  0.998537  0.007545
## scale(age, scale = FALSE):obeseYes            -0.007257  0.992769  0.008342
## scale(age, scale = FALSE):smokerCurrent       -0.024388  0.975907  0.008806
## scale(age, scale = FALSE):smokerFormer        -0.003120  0.996885  0.008877
## scale(age, scale = FALSE):strokeYes           -0.009273  0.990770  0.011277
##                                                    z Pr(>|z|)
## scale(age, scale = FALSE)                     10.856  < 2e-16 ***
## diabeticYes                                    5.288 1.24e-07 ***
## educ9th-11th grade                             0.040 0.968033
## educHS graduate                               -1.330 0.183538
## educSome college                              -1.645 0.100042
## educCollege graduate                          -4.394 1.11e-05 ***
## hypertensionYes                                3.712 0.000206 ***
## marital2Separated                              5.839 5.24e-09 ***
## marital2Single                                 4.227 2.37e-05 ***
## mi_chdYes                                      3.727 0.000194 ***
## underweightYes                                 3.084 0.002039 **
## overweight_exYes                              -0.792 0.428181
## obeseYes                                       0.551 0.581392
```

```
## smokerCurrent                                    7.671 1.71e-14 ***
## smokerFormer                                     1.798 0.072156 .
## strokeYes                                        3.973 7.10e-05 ***
## scale(age, scale = FALSE):diabeticYes           -3.125 0.001777 **
## scale(age, scale = FALSE):educ9th-11th grade     1.118 0.263559
## scale(age, scale = FALSE):educHS graduate        0.540 0.588928
## scale(age, scale = FALSE):educSome college       0.222 0.824594
## scale(age, scale = FALSE):educCollege graduate   1.369 0.170923
## scale(age, scale = FALSE):hypertensionYes       -3.193 0.001410 **
## scale(age, scale = FALSE):marital2Separated     -1.763 0.077944 .
## scale(age, scale = FALSE):marital2Single        -0.925 0.354912
## scale(age, scale = FALSE):mi_chdYes             -1.281 0.200020
## scale(age, scale = FALSE):underweightYes        -0.606 0.544336
## scale(age, scale = FALSE):overweight_exYes      -0.194 0.846189
## scale(age, scale = FALSE):obeseYes              -0.870 0.384328
## scale(age, scale = FALSE):smokerCurrent         -2.769 0.005615 **
## scale(age, scale = FALSE):smokerFormer          -0.352 0.725211
## scale(age, scale = FALSE):strokeYes             -0.822 0.410896
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                 exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                          1.1246     0.8892    1.1010
## diabeticYes                                        1.7705     0.5648    1.4327
## educ9th-11th grade                                 1.0046     0.9954    0.8026
## educHS graduate                                    0.8340     1.1991    0.6382
## educSome college                                   0.8101     1.2344    0.6303
## educCollege graduate                               0.5196     1.9244    0.3881
## hypertensionYes                                    1.3795     0.7249    1.1640
## marital2Separated                                  1.6756     0.5968    1.4091
## marital2Single                                     1.7953     0.5570    1.3686
## mi_chdYes                                          1.4839     0.6739    1.2058
## underweightYes                                     2.6254     0.3809    1.4218
## overweight_exYes                                   0.9136     1.0946    0.7306
## obeseYes                                           1.0681     0.9363    0.8451
## smokerCurrent                                      2.3599     0.4237    1.8951
## smokerFormer                                       1.2381     0.8077    0.9810
## strokeYes                                          1.8784     0.5324    1.3763
## scale(age, scale = FALSE):diabeticYes              0.9751     1.0255    0.9599
## scale(age, scale = FALSE):educ9th-11th grade       1.0100     0.9901    0.9925
## scale(age, scale = FALSE):educHS graduate          1.0048     0.9952    0.9875
## scale(age, scale = FALSE):educSome college         1.0018     0.9982    0.9858
## scale(age, scale = FALSE):educCollege graduate     1.0148     0.9855    0.9937
## scale(age, scale = FALSE):hypertensionYes          0.9807     1.0197    0.9690
## scale(age, scale = FALSE):marital2Separated        0.9879     1.0122    0.9747
## scale(age, scale = FALSE):marital2Single           0.9897     1.0104    0.9682
## scale(age, scale = FALSE):mi_chdYes                0.9898     1.0103    0.9745
## scale(age, scale = FALSE):underweightYes           0.9831     1.0172    0.9302
## scale(age, scale = FALSE):overweight_exYes         0.9985     1.0015    0.9839
## scale(age, scale = FALSE):obeseYes                 0.9928     1.0073    0.9767
## scale(age, scale = FALSE):smokerCurrent            0.9759     1.0247    0.9592
## scale(age, scale = FALSE):smokerFormer             0.9969     1.0031    0.9797
## scale(age, scale = FALSE):strokeYes                0.9908     1.0093    0.9691
```

```
##                                                 upper .95
## scale(age, scale = FALSE)                          1.1487
## diabeticYes                                        2.1880
## educ9th-11th grade                                 1.2575
## educHS graduate                                    1.0898
## educSome college                                   1.0412
## educCollege graduate                               0.6958
## hypertensionYes                                    1.6349
## marital2Separated                                  1.9926
## marital2Single                                     2.3549
## mi_chdYes                                          1.8263
## underweightYes                                     4.8481
## overweight_exYes                                   1.1424
## obeseYes                                           1.3499
## smokerCurrent                                      2.9388
## smokerFormer                                       1.5626
## strokeYes                                          2.5637
## scale(age, scale = FALSE):diabeticYes              0.9907
## scale(age, scale = FALSE):educ9th-11th grade       1.0278
## scale(age, scale = FALSE):educHS graduate          1.0224
## scale(age, scale = FALSE):educSome college         1.0181
## scale(age, scale = FALSE):educCollege graduate     1.0362
## scale(age, scale = FALSE):hypertensionYes          0.9925
## scale(age, scale = FALSE):marital2Separated        1.0014
## scale(age, scale = FALSE):marital2Single           1.0117
## scale(age, scale = FALSE):mi_chdYes                1.0054
## scale(age, scale = FALSE):underweightYes           1.0389
## scale(age, scale = FALSE):overweight_exYes         1.0134
## scale(age, scale = FALSE):obeseYes                 1.0091
## scale(age, scale = FALSE):smokerCurrent            0.9929
## scale(age, scale = FALSE):smokerFormer             1.0144
## scale(age, scale = FALSE):strokeYes                1.0129
##
## Concordance= 0.823  (se = 0.008 )
## Likelihood ratio test= NA  on 31 df,   p=NA
## Wald test            = 2326  on 31 df,   p=<2e-16
## Score (logrank) test = NA  on 31 df,   p=NA
```

As we might expect, broadening out to a larger sample reduced predictive performance somewhat (0.84 vs. 0.82). Introducing interactions boosted performance somewhat. Let's drop the interactions between age and marital status, MI/CHD, smoking status, and weight, since they seem less necessary:

```
plco_int2 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + mi_chd + underweight + overweight_ex + obese +
      smoker + stroke + scale(age, scale = FALSE) * diabetic +
      scale(age, scale = FALSE) * educ + scale(age, scale = FALSE) * hypertension +
      scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int2)
## Stratified 1 - level Cluster Sampling design (with replacement)
```

```
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + mi_chd + underweight +
##     overweight_ex + obese + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     stroke, design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                                       coef  exp(coef)   se(coef)
## scale(age, scale = FALSE)                          0.1028345  1.1083079  0.0077550
## diabeticYes                                        0.5692250  1.7668971  0.1058272
## educ9th-11th grade                                 0.0540365  1.0555231  0.1178452
## educHS graduate                                   -0.1570845  0.8546319  0.1414679
## educSome college                                  -0.1999785  0.8187484  0.1320024
## educCollege graduate                              -0.6890905  0.5020324  0.1489595
## hypertensionYes                                    0.3384376  1.4027542  0.0857826
## marital2Separated                                  0.4183046  1.5193834  0.0610329
## marital2Single                                     0.5714380  1.7708116  0.1407590
## mi_chdYes                                          0.3022378  1.3528829  0.0656896
## underweightYes                                     0.9044406  2.4705495  0.2893756
## overweight_exYes                                  -0.1067038  0.8987919  0.0746672
## obeseYes                                           0.0093637  1.0094077  0.0838898
## smokerCurrent                                      0.8083725  2.2442525  0.0909883
## smokerFormer                                       0.1948914  1.2151790  0.0647977
## strokeYes                                          0.6601520  1.9350865  0.1533825
## scale(age, scale = FALSE):diabeticYes             -0.0268932  0.9734652  0.0079263
## scale(age, scale = FALSE):educ9th-11th grade       0.0057624  1.0057791  0.0093770
## scale(age, scale = FALSE):educHS graduate          0.0031385  1.0031434  0.0090278
## scale(age, scale = FALSE):educSome college         0.0009502  1.0009507  0.0084634
## scale(age, scale = FALSE):educCollege graduate     0.0183189  1.0184878  0.0105164
## scale(age, scale = FALSE):hypertensionYes         -0.0213029  0.9789224  0.0059255
## scale(age, scale = FALSE):strokeYes               -0.0111738  0.9888884  0.0109213
##                                                       z Pr(>|z|)
## scale(age, scale = FALSE)                          13.260  < 2e-16 ***
## diabeticYes                                         5.379 7.50e-08 ***
## educ9th-11th grade                                  0.459 0.646566
## educHS graduate                                    -1.110 0.266831
## educSome college                                   -1.515 0.129782
## educCollege graduate                               -4.626 3.73e-06 ***
## hypertensionYes                                     3.945 7.97e-05 ***
## marital2Separated                                   6.854 7.19e-12 ***
## marital2Single                                      4.060 4.91e-05 ***
## mi_chdYes                                           4.601 4.20e-06 ***
## underweightYes                                      3.125 0.001775 **
## overweight_exYes                                   -1.429 0.152987
## obeseYes                                            0.112 0.911125
## smokerCurrent                                       8.884  < 2e-16 ***
## smokerFormer                                        3.008 0.002632 **
## strokeYes                                           4.304 1.68e-05 ***
```

```
## scale(age, scale = FALSE):diabeticYes          -3.393 0.000692 ***
## scale(age, scale = FALSE):educ9th-11th grade    0.615 0.538866
## scale(age, scale = FALSE):educHS graduate       0.348 0.728105
## scale(age, scale = FALSE):educSome college      0.112 0.910605
## scale(age, scale = FALSE):educCollege graduate  1.742 0.081519 .
## scale(age, scale = FALSE):hypertensionYes      -3.595 0.000324 ***
## scale(age, scale = FALSE):strokeYes            -1.023 0.306248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                         1.1083     0.9023    1.0916
## diabeticYes                                       1.7669     0.5660    1.4359
## educ9th-11th grade                                1.0555     0.9474    0.8378
## educHS graduate                                   0.8546     1.1701    0.6477
## educSome college                                  0.8187     1.2214    0.6321
## educCollege graduate                              0.5020     1.9919    0.3749
## hypertensionYes                                   1.4028     0.7129    1.1857
## marital2Separated                                 1.5194     0.6582    1.3481
## marital2Single                                    1.7708     0.5647    1.3439
## mi_chdYes                                         1.3529     0.7392    1.1894
## underweightYes                                    2.4705     0.4048    1.4011
## overweight_exYes                                  0.8988     1.1126    0.7764
## obeseYes                                          1.0094     0.9907    0.8564
## smokerCurrent                                     2.2443     0.4456    1.8777
## smokerFormer                                      1.2152     0.8229    1.0702
## strokeYes                                         1.9351     0.5168    1.4327
## scale(age, scale = FALSE):diabeticYes             0.9735     1.0273    0.9585
## scale(age, scale = FALSE):educ9th-11th grade      1.0058     0.9943    0.9875
## scale(age, scale = FALSE):educHS graduate         1.0031     0.9969    0.9855
## scale(age, scale = FALSE):educSome college        1.0010     0.9991    0.9845
## scale(age, scale = FALSE):educCollege graduate    1.0185     0.9818    0.9977
## scale(age, scale = FALSE):hypertensionYes         0.9789     1.0215    0.9676
## scale(age, scale = FALSE):strokeYes               0.9889     1.0112    0.9679
##                                                upper .95
## scale(age, scale = FALSE)                         1.1253
## diabeticYes                                       2.1742
## educ9th-11th grade                                1.3298
## educHS graduate                                   1.1277
## educSome college                                  1.0605
## educCollege graduate                              0.6722
## hypertensionYes                                   1.6596
## marital2Separated                                 1.7125
## marital2Single                                    2.3334
## mi_chdYes                                         1.5388
## underweightYes                                    4.3562
## overweight_exYes                                  1.0404
## obeseYes                                          1.1898
## smokerCurrent                                     2.6824
## smokerFormer                                      1.3797
## strokeYes                                         2.6137
## scale(age, scale = FALSE):diabeticYes             0.9887
## scale(age, scale = FALSE):educ9th-11th grade      1.0244
```

```
## scale(age, scale = FALSE):educHS graduate          1.0211
## scale(age, scale = FALSE):educSome college         1.0177
## scale(age, scale = FALSE):educCollege graduate     1.0397
## scale(age, scale = FALSE):hypertensionYes          0.9904
## scale(age, scale = FALSE):strokeYes                1.0103
##
## Concordance= 0.822  (se = 0.008 )
## Likelihood ratio test= NA  on 23 df,   p=NA
## Wald test            = 2257  on 23 df,   p=<2e-16
## Score (logrank) test = NA  on 23 df,   p=NA
```

We now try dropping each of the effects without interactions (marital status, MI/CHD, weight, smoking) without hurting model predictive performance substantially, in order to have a more parsimonious model. From this, it seems that we can drop MI/CHD without any major ill effects, so we do:

```
plco_int3 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight_ex + obese + smoker +
      stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int3)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + underweight +
##     overweight_ex + obese + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     stroke, design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                                    coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                      0.102958  1.108445  0.007720
## diabeticYes                                    0.591903  1.807424  0.106172
## educ9th-11th grade                             0.059497  1.061302  0.118335
## educHS graduate                               -0.163677  0.849016  0.140814
## educSome college                              -0.205078  0.814584  0.132497
## educCollege graduate                          -0.698738  0.497212  0.149171
## hypertensionYes                                0.362401  1.436774  0.084662
## marital2Separated                              0.420421  1.522603  0.060634
## marital2Single                                 0.560065  1.750787  0.141598
## underweightYes                                 0.883691  2.419814  0.290719
## overweight_exYes                              -0.114996  0.891370  0.072920
## obeseYes                                       0.008240  1.008274  0.083754
## smokerCurrent                                  0.820625  2.271919  0.089334
## smokerFormer                                   0.221977  1.248543  0.065781
## strokeYes                                      0.710190  2.034377  0.157515
```

```
## scale(age, scale = FALSE):diabeticYes          -0.026281  0.974061  0.007909
## scale(age, scale = FALSE):educ9th-11th grade    0.006428  1.006449  0.009187
## scale(age, scale = FALSE):educHS graduate       0.004635  1.004646  0.009065
## scale(age, scale = FALSE):educSome college      0.002123  1.002125  0.008546
## scale(age, scale = FALSE):educCollege graduate  0.020266  1.020473  0.010506
## scale(age, scale = FALSE):hypertensionYes      -0.020818  0.979397  0.005969
## scale(age, scale = FALSE):strokeYes            -0.011182  0.988880  0.011079
##                                                      z Pr(>|z|)
## scale(age, scale = FALSE)                       13.337  < 2e-16 ***
## diabeticYes                                      5.575 2.48e-08 ***
## educ9th-11th grade                               0.503 0.615118
## educHS graduate                                 -1.162 0.245089
## educSome college                                -1.548 0.121670
## educCollege graduate                            -4.684 2.81e-06 ***
## hypertensionYes                                  4.281 1.86e-05 ***
## marital2Separated                                6.934 4.10e-12 ***
## marital2Single                                   3.955 7.64e-05 ***
## underweightYes                                   3.040 0.002368 **
## overweight_exYes                                -1.577 0.114790
## obeseYes                                         0.098 0.921626
## smokerCurrent                                    9.186  < 2e-16 ***
## smokerFormer                                     3.375 0.000739 ***
## strokeYes                                        4.509 6.52e-06 ***
## scale(age, scale = FALSE):diabeticYes           -3.323 0.000891 ***
## scale(age, scale = FALSE):educ9th-11th grade     0.700 0.484109
## scale(age, scale = FALSE):educHS graduate        0.511 0.609125
## scale(age, scale = FALSE):educSome college       0.248 0.803825
## scale(age, scale = FALSE):educCollege graduate   1.929 0.053738 .
## scale(age, scale = FALSE):hypertensionYes       -3.488 0.000487 ***
## scale(age, scale = FALSE):strokeYes             -1.009 0.312834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                 exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                          1.1084     0.9022    1.0918
## diabeticYes                                        1.8074     0.5533    1.4679
## educ9th-11th grade                                 1.0613     0.9422    0.8416
## educHS graduate                                    0.8490     1.1778    0.6443
## educSome college                                   0.8146     1.2276    0.6283
## educCollege graduate                               0.4972     2.0112    0.3712
## hypertensionYes                                    1.4368     0.6960    1.2171
## marital2Separated                                  1.5226     0.6568    1.3520
## marital2Single                                     1.7508     0.5712    1.3265
## underweightYes                                     2.4198     0.4133    1.3687
## overweight_exYes                                   0.8914     1.1219    0.7727
## obeseYes                                           1.0083     0.9918    0.8556
## smokerCurrent                                      2.2719     0.4402    1.9070
## smokerFormer                                       1.2485     0.8009    1.0975
## strokeYes                                          2.0344     0.4916    1.4940
## scale(age, scale = FALSE):diabeticYes              0.9741     1.0266    0.9591
## scale(age, scale = FALSE):educ9th-11th grade       1.0064     0.9936    0.9885
## scale(age, scale = FALSE):educHS graduate          1.0046     0.9954    0.9870
## scale(age, scale = FALSE):educSome college         1.0021     0.9979    0.9855
```

```
## scale(age, scale = FALSE):educCollege graduate    1.0205     0.9799    0.9997
## scale(age, scale = FALSE):hypertensionYes         0.9794     1.0210    0.9680
## scale(age, scale = FALSE):strokeYes               0.9889     1.0112    0.9676
##                                                 upper .95
## scale(age, scale = FALSE)                          1.1253
## diabeticYes                                        2.2255
## educ9th-11th grade                                 1.3383
## educHS graduate                                    1.1189
## educSome college                                   1.0561
## educCollege graduate                               0.6661
## hypertensionYes                                    1.6961
## marital2Separated                                  1.7147
## marital2Single                                     2.3108
## underweightYes                                     4.2780
## overweight_exYes                                   1.0283
## obeseYes                                           1.1881
## smokerCurrent                                      2.7067
## smokerFormer                                       1.4204
## strokeYes                                          2.7702
## scale(age, scale = FALSE):diabeticYes              0.9893
## scale(age, scale = FALSE):educ9th-11th grade       1.0247
## scale(age, scale = FALSE):educHS graduate          1.0227
## scale(age, scale = FALSE):educSome college         1.0191
## scale(age, scale = FALSE):educCollege graduate     1.0417
## scale(age, scale = FALSE):hypertensionYes          0.9909
## scale(age, scale = FALSE):strokeYes                1.0106
##
## Concordance= 0.821  (se = 0.008 )
## Likelihood ratio test= NA  on 22 df,   p=NA
## Wald test            = 2305  on 22 df,   p=<2e-16
## Score (logrank) test = NA  on 22 df,   p=NA
```

All right. This seems like a pretty solid model. We will now consider adding in income information, which was a covariate of interest for our collaborators (but has substantial missingness, so we excluded it until now):

```
plco_int4 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight_ex + obese + smoker +
      stroke + INDFMPIR + insurance + scale(age, scale = FALSE) * diabetic +
      scale(age, scale = FALSE) * educ + scale(age, scale = FALSE) * hypertension +
      scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int4)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + underweight +
##     overweight_ex + obese + smoker + stroke + INDFMPIR + insurance +
##     scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) *
```

```
##     educ + scale(age, scale = FALSE) * hypertension + scale(age,
##     scale = FALSE) * stroke, design = nhanes_sub2)
##
##   n= 6747, number of events= 1613
##    (582 observations deleted due to missingness)
##
##                                                    coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                      0.103976  1.109573  0.008051
## diabeticYes                                    0.616114  1.851719  0.101390
## educ9th-11th grade                             0.159333  1.172729  0.132944
## educHS graduate                               -0.007364  0.992663  0.163174
## educSome college                              -0.068023  0.934239  0.147665
## educCollege graduate                          -0.427512  0.652130  0.180685
## hypertensionYes                                0.347542  1.415583  0.096976
## marital2Separated                              0.351667  1.421434  0.066379
## marital2Single                                 0.445536  1.561326  0.146062
## underweightYes                                 1.121485  3.069408  0.311301
## overweight_exYes                              -0.036443  0.964213  0.081562
## obeseYes                                       0.076877  1.079909  0.089685
## smokerCurrent                                  0.808592  2.244745  0.097635
## smokerFormer                                   0.246094  1.279020  0.066403
## strokeYes                                      0.688507  1.990742  0.176218
## INDFMPIR                                      -0.129753  0.878313  0.032982
## insuranceYes                                  -0.066110  0.936028  0.123625
## scale(age, scale = FALSE):diabeticYes         -0.026900  0.973458  0.007531
## scale(age, scale = FALSE):educ9th-11th grade   0.003977  1.003985  0.009893
## scale(age, scale = FALSE):educHS graduate      0.005812  1.005829  0.009833
## scale(age, scale = FALSE):educSome college     0.004419  1.004428  0.008731
## scale(age, scale = FALSE):educCollege graduate 0.021344  1.021573  0.010863
## scale(age, scale = FALSE):hypertensionYes     -0.021659  0.978574  0.006218
## scale(age, scale = FALSE):strokeYes           -0.013370  0.986719  0.012351
##                                                    z Pr(>|z|)
## scale(age, scale = FALSE)                     12.914  < 2e-16 ***
## diabeticYes                                    6.077 1.23e-09 ***
## educ9th-11th grade                             1.198 0.230724
## educHS graduate                               -0.045 0.964006
## educSome college                              -0.461 0.645042
## educCollege graduate                          -2.366 0.017978 *
## hypertensionYes                                3.584 0.000339 ***
## marital2Separated                              5.298 1.17e-07 ***
## marital2Single                                 3.050 0.002286 **
## underweightYes                                 3.603 0.000315 ***
## overweight_exYes                              -0.447 0.655008
## obeseYes                                       0.857 0.391341
## smokerCurrent                                  8.282  < 2e-16 ***
## smokerFormer                                   3.706 0.000210 ***
## strokeYes                                      3.907 9.34e-05 ***
## INDFMPIR                                      -3.934 8.35e-05 ***
## insuranceYes                                  -0.535 0.592817
## scale(age, scale = FALSE):diabeticYes         -3.572 0.000354 ***
## scale(age, scale = FALSE):educ9th-11th grade   0.402 0.687660
## scale(age, scale = FALSE):educHS graduate      0.591 0.554427
## scale(age, scale = FALSE):educSome college     0.506 0.612787
```

```
## scale(age, scale = FALSE):educCollege graduate  1.965 0.049441 *
## scale(age, scale = FALSE):hypertensionYes      -3.483 0.000495 ***
## scale(age, scale = FALSE):strokeYes            -1.082 0.279032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                         1.1096     0.9012    1.0922
## diabeticYes                                       1.8517     0.5400    1.5180
## educ9th-11th grade                                1.1727     0.8527    0.9037
## educHS graduate                                   0.9927     1.0074    0.7210
## educSome college                                  0.9342     1.0704    0.6995
## educCollege graduate                              0.6521     1.5334    0.4577
## hypertensionYes                                   1.4156     0.7064    1.1705
## marital2Separated                                 1.4214     0.7035    1.2480
## marital2Single                                    1.5613     0.6405    1.1726
## underweightYes                                    3.0694     0.3258    1.6675
## overweight_exYes                                  0.9642     1.0371    0.8218
## obeseYes                                          1.0799     0.9260    0.9058
## smokerCurrent                                     2.2447     0.4455    1.8538
## smokerFormer                                      1.2790     0.7818    1.1229
## strokeYes                                         1.9907     0.5023    1.4093
## INDFMPIR                                          0.8783     1.1385    0.8233
## insuranceYes                                      0.9360     1.0683    0.7346
## scale(age, scale = FALSE):diabeticYes             0.9735     1.0273    0.9592
## scale(age, scale = FALSE):educ9th-11th grade      1.0040     0.9960    0.9847
## scale(age, scale = FALSE):educHS graduate         1.0058     0.9942    0.9866
## scale(age, scale = FALSE):educSome college        1.0044     0.9956    0.9874
## scale(age, scale = FALSE):educCollege graduate    1.0216     0.9789    1.0001
## scale(age, scale = FALSE):hypertensionYes         0.9786     1.0219    0.9667
## scale(age, scale = FALSE):strokeYes               0.9867     1.0135    0.9631
##                                                upper .95
## scale(age, scale = FALSE)                         1.1272
## diabeticYes                                       2.2588
## educ9th-11th grade                                1.5218
## educHS graduate                                   1.3668
## educSome college                                  1.2478
## educCollege graduate                              0.9293
## hypertensionYes                                   1.7119
## marital2Separated                                 1.6189
## marital2Single                                    2.0788
## underweightYes                                    5.6498
## overweight_exYes                                  1.1314
## obeseYes                                          1.2874
## smokerCurrent                                     2.7182
## smokerFormer                                      1.4568
## strokeYes                                         2.8120
## INDFMPIR                                          0.9370
## insuranceYes                                      1.1927
## scale(age, scale = FALSE):diabeticYes             0.9879
## scale(age, scale = FALSE):educ9th-11th grade      1.0236
## scale(age, scale = FALSE):educHS graduate         1.0254
## scale(age, scale = FALSE):educSome college        1.0218
```

```
## scale(age, scale = FALSE):educCollege graduate    1.0436
## scale(age, scale = FALSE):hypertensionYes         0.9906
## scale(age, scale = FALSE):strokeYes               1.0109
##
## Concordance= 0.83  (se = 0.008 )
## Likelihood ratio test= NA  on 24 df,   p=NA
## Wald test            = 2407  on 24 df,   p=<2e-16
## Score (logrank) test = NA  on 24 df,   p=NA
```

That boosted performance somewhat (0.82 vs. 0.83), but not enough to deal with the hassle of imputing it in PLCO, which doesn't have any income information. Let's revert to our previous model.

My collaborators were concerned by the direction of the effects of being overweight (currently protective). We try a continuous version of BMI, both with and without a quadratic term:

```
plco_int5 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + BMXBMI + smoker + stroke +
      scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int5)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + BMXBMI + smoker +
##     stroke + scale(age, scale = FALSE) * diabetic + scale(age,
##     scale = FALSE) * educ + scale(age, scale = FALSE) * hypertension +
##     scale(age, scale = FALSE) * stroke, design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                                  coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                    0.104189  1.109810  0.007825
## diabeticYes                                  0.548461  1.730587  0.108450
## educ9th-11th grade                           0.069249  1.071703  0.118193
## educHS graduate                             -0.173286  0.840897  0.140215
## educSome college                            -0.213482  0.807766  0.129723
## educCollege graduate                        -0.692860  0.500144  0.150174
## hypertensionYes                              0.334061  1.396629  0.085195
## marital2Separated                            0.432584  1.541235  0.060493
## marital2Single                               0.628441  1.874686  0.137993
## BMXBMI                                       0.013470  1.013561  0.006978
## smokerCurrent                                0.864643  2.374159  0.089884
## smokerFormer                                 0.214786  1.239596  0.064744
## strokeYes                                    0.709283  2.032533  0.154230
## scale(age, scale = FALSE):diabeticYes       -0.025972  0.974363  0.007868
## scale(age, scale = FALSE):educ9th-11th grade 0.006988  1.007013  0.009160
## scale(age, scale = FALSE):educHS graduate    0.004659  1.004669  0.009046
## scale(age, scale = FALSE):educSome college   0.002254  1.002257  0.008397
```

```
## scale(age, scale = FALSE):educCollege graduate  0.019890  1.020089  0.010538
## scale(age, scale = FALSE):hypertensionYes      -0.020656  0.979556  0.006009
## scale(age, scale = FALSE):strokeYes            -0.011667  0.988401  0.010795
##                                                      z Pr(>|z|)
## scale(age, scale = FALSE)                       13.314  < 2e-16 ***
## diabeticYes                                      5.057 4.25e-07 ***
## educ9th-11th grade                               0.586 0.557941
## educHS graduate                                 -1.236 0.216512
## educSome college                                -1.646 0.099830 .
## educCollege graduate                            -4.614 3.96e-06 ***
## hypertensionYes                                  3.921 8.81e-05 ***
## marital2Separated                                7.151 8.62e-13 ***
## marital2Single                                   4.554 5.26e-06 ***
## BMXBMI                                           1.930 0.053557 .
## smokerCurrent                                    9.620  < 2e-16 ***
## smokerFormer                                     3.317 0.000908 ***
## strokeYes                                        4.599 4.25e-06 ***
## scale(age, scale = FALSE):diabeticYes           -3.301 0.000964 ***
## scale(age, scale = FALSE):educ9th-11th grade     0.763 0.445520
## scale(age, scale = FALSE):educHS graduate        0.515 0.606574
## scale(age, scale = FALSE):educSome college       0.268 0.788345
## scale(age, scale = FALSE):educCollege graduate   1.887 0.059109 .
## scale(age, scale = FALSE):hypertensionYes       -3.438 0.000587 ***
## scale(age, scale = FALSE):strokeYes             -1.081 0.279816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                 exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                          1.1098     0.9011    1.0929
## diabeticYes                                        1.7306     0.5778    1.3992
## educ9th-11th grade                                 1.0717     0.9331    0.8501
## educHS graduate                                    0.8409     1.1892    0.6388
## educSome college                                   0.8078     1.2380    0.6264
## educCollege graduate                               0.5001     1.9994    0.3726
## hypertensionYes                                    1.3966     0.7160    1.1819
## marital2Separated                                  1.5412     0.6488    1.3689
## marital2Single                                     1.8747     0.5334    1.4304
## BMXBMI                                             1.0136     0.9866    0.9998
## smokerCurrent                                      2.3742     0.4212    1.9907
## smokerFormer                                       1.2396     0.8067    1.0919
## strokeYes                                          2.0325     0.4920    1.5023
## scale(age, scale = FALSE):diabeticYes              0.9744     1.0263    0.9595
## scale(age, scale = FALSE):educ9th-11th grade       1.0070     0.9930    0.9891
## scale(age, scale = FALSE):educHS graduate          1.0047     0.9954    0.9870
## scale(age, scale = FALSE):educSome college         1.0023     0.9977    0.9859
## scale(age, scale = FALSE):educCollege graduate     1.0201     0.9803    0.9992
## scale(age, scale = FALSE):hypertensionYes          0.9796     1.0209    0.9681
## scale(age, scale = FALSE):strokeYes                0.9884     1.0117    0.9677
##                                                 upper .95
## scale(age, scale = FALSE)                          1.1270
## diabeticYes                                        2.1405
## educ9th-11th grade                                 1.3511
## educHS graduate                                    1.1069
```

```
## educSome college                                    1.0416
## educCollege graduate                                0.6713
## hypertensionYes                                     1.6504
## marital2Separated                                   1.7352
## marital2Single                                      2.4569
## BMXBMI                                              1.0275
## smokerCurrent                                       2.8315
## smokerFormer                                        1.4073
## strokeYes                                           2.7499
## scale(age, scale = FALSE):diabeticYes               0.9895
## scale(age, scale = FALSE):educ9th-11th grade        1.0253
## scale(age, scale = FALSE):educHS graduate           1.0226
## scale(age, scale = FALSE):educSome college          1.0189
## scale(age, scale = FALSE):educCollege graduate      1.0414
## scale(age, scale = FALSE):hypertensionYes           0.9912
## scale(age, scale = FALSE):strokeYes                 1.0095
##
## Concordance= 0.819  (se = 0.008 )
## Likelihood ratio test= NA  on 20 df,   p=NA
## Wald test            = 2264  on 20 df,   p=<2e-16
## Score (logrank) test = NA  on 20 df,   p=NA

mydata$bmi_sq <- mydata$BMXBMI * mydata$BMXBMI
nhanes_svy <-
  svydesign(
    data = mydata,
    id =  ~ psu,
    strata =  ~ strata,
    weights =  ~ sweights,
    nest = TRUE
  )
nhanes_sub2 <- subset(nhanes_svy, inmodel2 == 1)

plco_int6 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + BMXBMI + bmi_sq + smoker + stroke +
      scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
    design = nhanes_sub2
  )

summary(plco_int6)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + BMXBMI + bmi_sq +
##     smoker + stroke + scale(age, scale = FALSE) * diabetic +
##     scale(age, scale = FALSE) * educ + scale(age, scale = FALSE) *
##     hypertension + scale(age, scale = FALSE) * stroke, design = nhanes_sub2)
##
```

```
##   n= 7329, number of events= 1769
##
##                                                      coef  exp(coef)   se(coef)
## scale(age, scale = FALSE)                         0.1037588  1.1093329  0.0077869
## diabeticYes                                       0.5633296  1.7565112  0.1084644
## educ9th-11th grade                                0.0596267  1.0614403  0.1188126
## educHS graduate                                  -0.1704671  0.8432708  0.1402400
## educSome college                                 -0.2112962  0.8095342  0.1300634
## educCollege graduate                             -0.6919486  0.5005997  0.1501784
## hypertensionYes                                   0.3379408  1.4020576  0.0853068
## marital2Separated                                 0.4297947  1.5369420  0.0602696
## marital2Single                                    0.6076708  1.8361496  0.1393438
## BMXBMI                                           -0.0035559  0.9964504  0.0124625
## bmi_sq                                            0.0002115  1.0002115  0.0001156
## smokerCurrent                                     0.8634438  2.3713129  0.0898765
## smokerFormer                                      0.2203461  1.2465081  0.0651165
## strokeYes                                         0.7144292  2.0430201  0.1542576
## scale(age, scale = FALSE):diabeticYes            -0.0262315  0.9741096  0.0079018
## scale(age, scale = FALSE):educ9th-11th grade      0.0077782  1.0078086  0.0092484
## scale(age, scale = FALSE):educHS graduate         0.0046461  1.0046569  0.0090456
## scale(age, scale = FALSE):educSome college        0.0022820  1.0022846  0.0083943
## scale(age, scale = FALSE):educCollege graduate    0.0199695  1.0201703  0.0104965
## scale(age, scale = FALSE):hypertensionYes        -0.0204368  0.9797706  0.0059994
## scale(age, scale = FALSE):strokeYes              -0.0119213  0.9881494  0.0107766
##                                                        z Pr(>|z|)
## scale(age, scale = FALSE)                         13.325  < 2e-16 ***
## diabeticYes                                        5.194 2.06e-07 ***
## educ9th-11th grade                                 0.502 0.615769
## educHS graduate                                   -1.216 0.224161
## educSome college                                  -1.625 0.104256
## educCollege graduate                              -4.608 4.08e-06 ***
## hypertensionYes                                    3.961 7.45e-05 ***
## marital2Separated                                  7.131 9.95e-13 ***
## marital2Single                                     4.361 1.30e-05 ***
## BMXBMI                                            -0.285 0.775395
## bmi_sq                                             1.830 0.067238 .
## smokerCurrent                                      9.607  < 2e-16 ***
## smokerFormer                                       3.384 0.000715 ***
## strokeYes                                          4.631 3.63e-06 ***
## scale(age, scale = FALSE):diabeticYes             -3.320 0.000901 ***
## scale(age, scale = FALSE):educ9th-11th grade       0.841 0.400328
## scale(age, scale = FALSE):educHS graduate          0.514 0.607511
## scale(age, scale = FALSE):educSome college         0.272 0.785733
## scale(age, scale = FALSE):educCollege graduate     1.902 0.057107 .
## scale(age, scale = FALSE):hypertensionYes         -3.406 0.000658 ***
## scale(age, scale = FALSE):strokeYes               -1.106 0.268627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                  exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                           1.1093     0.9014    1.0925
## diabeticYes                                         1.7565     0.5693    1.4201
## educ9th-11th grade                                  1.0614     0.9421    0.8409
```

```
## educHS graduate                                      0.8433     1.1859    0.6406
## educSome college                                     0.8095     1.2353    0.6274
## educCollege graduate                                 0.5006     1.9976    0.3730
## hypertensionYes                                      1.4021     0.7132    1.1862
## marital2Separated                                    1.5369     0.6506    1.3657
## marital2Single                                       1.8361     0.5446    1.3973
## BMXBMI                                               0.9965     1.0036    0.9724
## bmi_sq                                               1.0002     0.9998    1.0000
## smokerCurrent                                        2.3713     0.4217    1.9883
## smokerFormer                                         1.2465     0.8022    1.0972
## strokeYes                                            2.0430     0.4895    1.5100
## scale(age, scale = FALSE):diabeticYes                0.9741     1.0266    0.9591
## scale(age, scale = FALSE):educ9th-11th grade         1.0078     0.9923    0.9897
## scale(age, scale = FALSE):educHS graduate            1.0047     0.9954    0.9870
## scale(age, scale = FALSE):educSome college           1.0023     0.9977    0.9859
## scale(age, scale = FALSE):educCollege graduate       1.0202     0.9802    0.9994
## scale(age, scale = FALSE):hypertensionYes            0.9798     1.0206    0.9683
## scale(age, scale = FALSE):strokeYes                  0.9881     1.0120    0.9675
##                                                    upper .95
## scale(age, scale = FALSE)                             1.1264
## diabeticYes                                           2.1726
## educ9th-11th grade                                    1.3398
## educHS graduate                                       1.1100
## educSome college                                      1.0446
## educCollege graduate                                  0.6719
## hypertensionYes                                       1.6572
## marital2Separated                                     1.7297
## marital2Single                                        2.4128
## BMXBMI                                                1.0211
## bmi_sq                                                1.0004
## smokerCurrent                                         2.8281
## smokerFormer                                          1.4162
## strokeYes                                             2.7642
## scale(age, scale = FALSE):diabeticYes                 0.9893
## scale(age, scale = FALSE):educ9th-11th grade          1.0262
## scale(age, scale = FALSE):educHS graduate             1.0226
## scale(age, scale = FALSE):educSome college            1.0189
## scale(age, scale = FALSE):educCollege graduate        1.0414
## scale(age, scale = FALSE):hypertensionYes             0.9914
## scale(age, scale = FALSE):strokeYes                   1.0092
##
## Concordance= 0.819  (se = 0.008 )
## Likelihood ratio test= NA  on 21 df,   p=NA
## Wald test            = 2259  on 21 df,   p=<2e-16
## Score (logrank) test = NA  on 21 df,   p=NA
```

We also try splitting overweight/obese at a BMI of 40:

```
plco_int7 <-
  svycoxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker +
      stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
```

```
    design = nhanes_sub2
  )

summary(plco_int7)
## Stratified 1 - level Cluster Sampling design (with replacement)
## With (180) clusters.
## subset(nhanes_svy, inmodel2 == 1)
## Call:
## svycoxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + underweight +
##     overweight2 + obese2 + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     stroke, design = nhanes_sub2)
##
##   n= 7329, number of events= 1769
##
##                                                      coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                        0.102767  1.108233  0.007694
## diabeticYes                                      0.561109  1.752615  0.107755
## educ9th-11th grade                               0.052202  1.053588  0.119104
## educHS graduate                                 -0.175748  0.838829  0.140065
## educSome college                                -0.207336  0.812747  0.131658
## educCollege graduate                            -0.686002  0.503585  0.149659
## hypertensionYes                                  0.352798  1.423043  0.085093
## marital2Separated                                0.420865  1.523279  0.060012
## marital2Single                                   0.579578  1.785284  0.139652
## underweightYes                                   0.882084  2.415929  0.290587
## overweight2Yes                                  -0.091325  0.912721  0.065632
## obese2Yes                                        0.590630  1.805124  0.173341
## smokerCurrent                                    0.831463  2.296677  0.089746
## smokerFormer                                     0.240157  1.271449  0.065342
## strokeYes                                        0.694475  2.002658  0.158654
## scale(age, scale = FALSE):diabeticYes           -0.024511  0.975787  0.007861
## scale(age, scale = FALSE):educ9th-11th grade     0.006538  1.006559  0.009233
## scale(age, scale = FALSE):educHS graduate        0.005264  1.005278  0.009125
## scale(age, scale = FALSE):educSome college       0.002582  1.002585  0.008458
## scale(age, scale = FALSE):educCollege graduate   0.020441  1.020651  0.010335
## scale(age, scale = FALSE):hypertensionYes       -0.020542  0.979667  0.005918
## scale(age, scale = FALSE):strokeYes             -0.009718  0.990329  0.011128
##                                                      z Pr(>|z|)
## scale(age, scale = FALSE)                       13.357  < 2e-16 ***
## diabeticYes                                      5.207 1.92e-07 ***
## educ9th-11th grade                               0.438 0.661177
## educHS graduate                                 -1.255 0.209567
## educSome college                                -1.575 0.115301
## educCollege graduate                            -4.584 4.57e-06 ***
## hypertensionYes                                  4.146 3.38e-05 ***
## marital2Separated                                7.013 2.33e-12 ***
## marital2Single                                   4.150 3.32e-05 ***
## underweightYes                                   3.036 0.002401 **
## overweight2Yes                                  -1.391 0.164085
## obese2Yes                                        3.407 0.000656 ***
```

```
## smokerCurrent                                      9.265  < 2e-16 ***
## smokerFormer                                       3.675 0.000238 ***
## strokeYes                                          4.377 1.20e-05 ***
## scale(age, scale = FALSE):diabeticYes             -3.118 0.001820 **
## scale(age, scale = FALSE):educ9th-11th grade       0.708 0.478893
## scale(age, scale = FALSE):educHS graduate          0.577 0.564010
## scale(age, scale = FALSE):educSome college         0.305 0.760143
## scale(age, scale = FALSE):educCollege graduate     1.978 0.047941 *
## scale(age, scale = FALSE):hypertensionYes         -3.471 0.000519 ***
## scale(age, scale = FALSE):strokeYes               -0.873 0.382501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                  exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                           1.1082     0.9023    1.0916
## diabeticYes                                         1.7526     0.5706    1.4189
## educ9th-11th grade                                  1.0536     0.9491    0.8342
## educHS graduate                                     0.8388     1.1921    0.6375
## educSome college                                    0.8127     1.2304    0.6279
## educCollege graduate                                0.5036     1.9858    0.3756
## hypertensionYes                                     1.4230     0.7027    1.2044
## marital2Separated                                   1.5233     0.6565    1.3542
## marital2Single                                      1.7853     0.5601    1.3578
## underweightYes                                      2.4159     0.4139    1.3669
## overweight2Yes                                      0.9127     1.0956    0.8026
## obese2Yes                                           1.8051     0.5540    1.2852
## smokerCurrent                                       2.2967     0.4354    1.9262
## smokerFormer                                        1.2714     0.7865    1.1186
## strokeYes                                           2.0027     0.4993    1.4674
## scale(age, scale = FALSE):diabeticYes               0.9758     1.0248    0.9609
## scale(age, scale = FALSE):educ9th-11th grade        1.0066     0.9935    0.9885
## scale(age, scale = FALSE):educHS graduate           1.0053     0.9947    0.9875
## scale(age, scale = FALSE):educSome college          1.0026     0.9974    0.9861
## scale(age, scale = FALSE):educCollege graduate      1.0207     0.9798    1.0002
## scale(age, scale = FALSE):hypertensionYes           0.9797     1.0208    0.9684
## scale(age, scale = FALSE):strokeYes                 0.9903     1.0098    0.9690
##                                                  upper .95
## scale(age, scale = FALSE)                           1.1251
## diabeticYes                                         2.1648
## educ9th-11th grade                                  1.3306
## educHS graduate                                     1.1038
## educSome college                                    1.0520
## educCollege graduate                                0.6752
## hypertensionYes                                     1.6813
## marital2Separated                                   1.7134
## marital2Single                                      2.3474
## underweightYes                                      4.2700
## overweight2Yes                                      1.0380
## obese2Yes                                           2.5355
## smokerCurrent                                       2.7384
## smokerFormer                                        1.4452
## strokeYes                                           2.7331
## scale(age, scale = FALSE):diabeticYes               0.9909
```

```
## scale(age, scale = FALSE):educ9th-11th grade      1.0249
## scale(age, scale = FALSE):educHS graduate         1.0234
## scale(age, scale = FALSE):educSome college        1.0193
## scale(age, scale = FALSE):educCollege graduate    1.0415
## scale(age, scale = FALSE):hypertensionYes         0.9911
## scale(age, scale = FALSE):strokeYes               1.0122
##
## Concordance= 0.822  (se = 0.008 )
## Likelihood ratio test= NA  on 22 df,   p=NA
## Wald test            = 2334  on 22 df,   p=<2e-16
## Score (logrank) test = NA  on 22 df,   p=NA
```

The split at 40 seems to be most interpretable without sacrificing predictive power, so we'll stick with that. We add in prostate cancer as a predictor because we'd like to adjust for that. We now refit this without the survey weights, so our final Cox model is:

```
mydata$inmodel4 <-
  ifelse(
    mydata$sex == "Male" &
      mydata$age > 40 &
      (mydata$cancer == 0 |
         mydata$pc == "PC") &
      !is.na(mydata$age) &
      !is.na(mydata$diabetic) &
      !is.na(mydata$educ) &
      !is.na(mydata$hypertension) &
      !is.na(mydata$marital2) &
      !is.na(mydata$BMXBMI) &
      !is.na(mydata$smoker) &
      !is.na(mydata$stroke) &
      !is.na(mydata$mortstat) & !is.na(mydata$pc),
    1,
    0
  )

cox_final <-
  coxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker +
      stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke + pc,
    data = mydata[mydata$inmodel4 == 1, ]
  )

summary(cox_final)
## Call:
## coxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + underweight +
##     overweight2 + obese2 + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     stroke + pc, data = mydata[mydata$inmodel4 == 1, ])
##
##   n= 7369, number of events= 1789
```

```
##
##                                                coef  exp(coef)   se(coef)
## scale(age, scale = FALSE)                  1.006e-01  1.106e+00  5.868e-03
## diabeticYes                                4.222e-01  1.525e+00  7.384e-02
## educ9th-11th grade                         2.701e-02  1.027e+00  1.021e-01
## educHS graduate                           -1.071e-01  8.984e-01  9.955e-02
## educSome college                          -1.211e-01  8.859e-01  9.940e-02
## educCollege graduate                      -5.710e-01  5.650e-01  1.252e-01
## hypertensionYes                            3.783e-01  1.460e+00  6.882e-02
## marital2Separated                          3.546e-01  1.426e+00  5.467e-02
## marital2Single                             5.702e-01  1.769e+00  1.062e-01
## underweightYes                             8.113e-01  2.251e+00  1.855e-01
## overweight2Yes                            -1.614e-01  8.510e-01  5.580e-02
## obese2Yes                                  3.246e-01  1.383e+00  1.354e-01
## smokerCurrent                              6.485e-01  1.913e+00  7.131e-02
## smokerFormer                               1.956e-01  1.216e+00  5.583e-02
## strokeYes                                  6.804e-01  1.975e+00  1.205e-01
## pcPC                                       1.083e-01  1.114e+00  8.232e-02
## scale(age, scale = FALSE):diabeticYes     -1.478e-02  9.853e-01  5.442e-03
## scale(age, scale = FALSE):educ9th-11th grade  4.844e-03  1.005e+00  7.151e-03
## scale(age, scale = FALSE):educHS graduate     5.429e-03  1.005e+00  6.837e-03
## scale(age, scale = FALSE):educSome college    6.849e-05  1.000e+00  6.942e-03
## scale(age, scale = FALSE):educCollege graduate  1.697e-02  1.017e+00  8.212e-03
## scale(age, scale = FALSE):hypertensionYes    -2.199e-02  9.782e-01  4.698e-03
## scale(age, scale = FALSE):strokeYes          -1.629e-02  9.838e-01  8.041e-03
##                                                 z Pr(>|z|)
## scale(age, scale = FALSE)                      17.148  < 2e-16 ***
## diabeticYes                                     5.717 1.08e-08 ***
## educ9th-11th grade                              0.264  0.79149
## educHS graduate                                -1.076  0.28179
## educSome college                               -1.219  0.22291
## educCollege graduate                           -4.561 5.10e-06 ***
## hypertensionYes                                 5.497 3.87e-08 ***
## marital2Separated                               6.486 8.81e-11 ***
## marital2Single                                  5.368 7.98e-08 ***
## underweightYes                                  4.373 1.23e-05 ***
## overweight2Yes                                 -2.892  0.00383 **
## obese2Yes                                       2.398  0.01650 *
## smokerCurrent                                   9.095  < 2e-16 ***
## smokerFormer                                    3.503  0.00046 ***
## strokeYes                                       5.648 1.62e-08 ***
## pcPC                                            1.316  0.18829
## scale(age, scale = FALSE):diabeticYes          -2.715  0.00662 **
## scale(age, scale = FALSE):educ9th-11th grade    0.677  0.49813
## scale(age, scale = FALSE):educHS graduate       0.794  0.42718
## scale(age, scale = FALSE):educSome college      0.010  0.99213
## scale(age, scale = FALSE):educCollege graduate  2.066  0.03883 *
## scale(age, scale = FALSE):hypertensionYes      -4.681 2.85e-06 ***
## scale(age, scale = FALSE):strokeYes            -2.025  0.04284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                            exp(coef) exp(-coef) lower .95
```

```
## scale(age, scale = FALSE)                          1.1059     0.9043    1.0932
## diabeticYes                                        1.5253     0.6556    1.3198
## educ9th-11th grade                                 1.0274     0.9734    0.8410
## educHS graduate                                    0.8984     1.1131    0.7391
## educSome college                                   0.8859     1.1288    0.7291
## educCollege graduate                               0.5650     1.7700    0.4420
## hypertensionYes                                    1.4598     0.6850    1.2756
## marital2Separated                                  1.4256     0.7015    1.2807
## marital2Single                                     1.7687     0.5654    1.4362
## underweightYes                                     2.2508     0.4443    1.5646
## overweight2Yes                                     0.8510     1.1751    0.7628
## obese2Yes                                          1.3835     0.7228    1.0611
## smokerCurrent                                      1.9127     0.5228    1.6632
## smokerFormer                                       1.2160     0.8224    1.0900
## strokeYes                                          1.9746     0.5064    1.5594
## pcPC                                               1.1144     0.8974    0.9483
## scale(age, scale = FALSE):diabeticYes              0.9853     1.0149    0.9749
## scale(age, scale = FALSE):educ9th-11th grade       1.0049     0.9952    0.9909
## scale(age, scale = FALSE):educHS graduate          1.0054     0.9946    0.9921
## scale(age, scale = FALSE):educSome college         1.0001     0.9999    0.9866
## scale(age, scale = FALSE):educCollege graduate     1.0171     0.9832    1.0009
## scale(age, scale = FALSE):hypertensionYes          0.9782     1.0222    0.9693
## scale(age, scale = FALSE):strokeYes                0.9838     1.0164    0.9685
##                                                  upper .95
## scale(age, scale = FALSE)                          1.1187
## diabeticYes                                        1.7628
## educ9th-11th grade                                 1.2551
## educHS graduate                                    1.0920
## educSome college                                   1.0764
## educCollege graduate                               0.7221
## hypertensionYes                                    1.6706
## marital2Separated                                  1.5868
## marital2Single                                     2.1781
## underweightYes                                     3.2378
## overweight2Yes                                     0.9493
## obese2Yes                                          1.8038
## smokerCurrent                                      2.1996
## smokerFormer                                       1.3566
## strokeYes                                          2.5004
## pcPC                                               1.3095
## scale(age, scale = FALSE):diabeticYes              0.9959
## scale(age, scale = FALSE):educ9th-11th grade       1.0190
## scale(age, scale = FALSE):educHS graduate          1.0190
## scale(age, scale = FALSE):educSome college         1.0138
## scale(age, scale = FALSE):educCollege graduate     1.0336
## scale(age, scale = FALSE):hypertensionYes          0.9873
## scale(age, scale = FALSE):strokeYes                0.9995
##
## Concordance= 0.796  (se = 0.005 )
## Likelihood ratio test= 2143  on 23 df,   p=<2e-16
## Wald test            = 1661  on 23 df,   p=<2e-16
## Score (logrank) test = 2202  on 23 df,   p=<2e-16
```

Random Forest Model Building

We also considered a survival random forest. First, we fit the random forest with all predictors:

```
mydata_nf$single <-
  ifelse(mydata_nf$marital2 == "Single", 1, ifelse(is.na(mydata_nf$marital2), NA_real_, 0))
mydata_nf$sep <-
  ifelse(mydata_nf$marital2 == "Separated",
         1,
         ifelse(is.na(mydata_nf$marital2), NA_real_, 0))
mydata_nf$black <-
  ifelse(mydata_nf$race2 == "NHB", 1, ifelse(is.na(mydata_nf$race2), NA_real_, 0))
mydata_nf$other <-
  ifelse(mydata_nf$race2 == "Other", 1, ifelse(is.na(mydata_nf$race2), NA_real_, 0))

rforest_nhanes <-
  rfsrc(
    Surv(permth_exm, mortstat) ~ age + alcoholic + anemic + arthritis + asthma +
      bronch + diabetic + educ + emphysema + hypertension + single + sep + mi +
      liver + black + other + smoker + stroke + underweight + overweight_ex +
      obese + angina + chf + chd + care + hospital + highchol + insurance +
      kidney + military + pc,
    data = mydata_nf[mydata_nf$inmodel1 == 1, ],
    seed = 1259
  )

varlist_nhanes <- vimp(rforest_nhanes)

print(rforest_nhanes)
##                          Sample size: 2420
##                     Number of deaths: 459
##                      Number of trees: 1000
##            Forest terminal node size: 15
##        Average no. of terminal nodes: 98.377
## No. of variables tried at each split: 6
##               Total no. of variables: 31
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 1529
##                             Analysis: RSF
##                               Family: surv
##                     Splitting rule: logrank *random*
##        Number of random split points: 10
##                           Error rate: 20.59%
print(varlist_nhanes$importance)
##           age     alcoholic        anemic     arthritis        asthma
##  1.270687e-01 -2.309169e-04  4.331114e-04 -2.860743e-05 -2.166845e-04
##        bronch      diabetic          educ     emphysema  hypertension
##  1.844676e-04  8.144145e-04  7.154104e-03  2.843074e-03  1.744025e-03
##        single           sep            mi         liver         black
##  2.061113e-03  6.384796e-03  1.579471e-03 -2.084481e-04 -4.227042e-04
##         other        smoker        stroke   underweight overweight_ex
##  1.386451e-04  4.741573e-03  4.414211e-03  1.169429e-03  9.724647e-04
##         obese        angina           chf           chd          care
##  2.667713e-04  5.512774e-04  5.724626e-03  3.772998e-04 -2.979062e-04
```

```
##      hospital      highchol     insurance        kidney      military
##  1.612249e-03  3.288005e-04  2.965866e-04  4.914734e-03  5.410574e-04
##            pc
##  3.472662e-03
```

And now we refit considering only PLCO-NHANES common covariates:

```
mydata_nf$inmodel3 <-
  ifelse(
    mydata_nf$sex == 1 &
      mydata_nf$age > 40 &
      (mydata_nf$cancer == 0 |
         mydata_nf$pc == 1) &
      !is.na(mydata_nf$age) &
      !is.na(mydata_nf$arthritis) &
      !is.na(mydata_nf$bronch) &
      !is.na(mydata_nf$diabetic) &
      !is.na(mydata_nf$educ) &
      !is.na(mydata_nf$emphysema) &
      !is.na(mydata_nf$hypertension) &
      !is.na(mydata_nf$marital2) &
      !is.na(mydata_nf$mi_chd) &
      !is.na(mydata_nf$BMXBMI) &
      !is.na(mydata_nf$liver) &
      !is.na(mydata_nf$race2) &
      !is.na(mydata_nf$smoker) &
      !is.na(mydata_nf$stroke) &
      !is.na(mydata_nf$mortstat) & !is.na(mydata_nf$pc),
    1,
    0
  )

rforest_plco <-
  rfsrc(
    Surv(permth_exm, mortstat) ~ age + arthritis + bronch + diabetic + educ +
      emphysema + hypertension + single + sep + mi_chd + underweight + overweight_ex +
      obese + liver + black + other + smoker + stroke + pc,
    data = mydata_nf[mydata_nf$inmodel3 == 1, ],
    seed = 1259
  )

varlist_plco <- vimp(rforest_plco)

print(rforest_plco)
##                          Sample size: 7268
##                     Number of deaths: 1751
##                      Number of trees: 1000
##            Forest terminal node size: 15
##        Average no. of terminal nodes: 283.749
## No. of variables tried at each split: 5
##               Total no. of variables: 19
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 4593
##                             Analysis: RSF
```

```
##                                Family: surv
##                     Splitting rule: logrank *random*
##        Number of random split points: 10
##                            Error rate: 20.87%
print(varlist_plco$importance)
##           age     arthritis        bronch      diabetic          educ
##  1.591383e-01 -1.808501e-04  1.214286e-04  3.967511e-03  4.637978e-03
##     emphysema  hypertension        single           sep        mi_chd
##  5.219042e-03  3.590087e-03  1.969771e-03  3.525419e-03  3.856568e-03
##   underweight overweight_ex         obese         liver         black
##  7.233762e-04  1.270915e-03  5.400237e-05 -1.166856e-04  3.696738e-04
##         other        smoker        stroke            pc
##  1.286139e-03  4.633181e-03  5.911927e-03  3.493949e-03
```

Again, model performance seems roughly comparable, so we will stick with the PLCO-NHANES common model. We do not need to consider interactions or other covariates, so the PLCO-NHANES common model is our final random forest candidate.

Parametric Spline Model Building

We consider refitting our original Cox model using parametric spline modeling. Note that we had to exclude 3 patients who had survival time of 0 in order to get these models to fit. We consider models with one vs. two knots, and time dependency on age:

```
sp1 <-
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker +
      stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke + pc,
    data = mydata[mydata$inmodel4 == 1 &
                    mydata$permth_exm > 0, ],
    k = 1,
    scale = "hazard"
  )

coefficients(sp1)
##                                      gamma0
##                                -8.1874060846
##                                      gamma1
##                                 1.1159895592
##                                      gamma2
##                                -0.0310416785
##                     scale(age, scale = FALSE)
##                                 0.0996964775
##                                  diabeticYes
##                                 0.4206839034
##                            educ9th-11th grade
##                                 0.0255687251
##                              educHS graduate
##                                -0.1115708111
##                            educSome college
##                                -0.1257730570
```

```
##                              educCollege graduate
##                                      -0.5764592131
##                                   hypertensionYes
##                                       0.3760238432
##                                 marital2Separated
##                                       0.3521364377
##                                    marital2Single
##                                       0.5667523573
##                                   underweightYes
##                                       0.7744055140
##                                   overweight2Yes
##                                      -0.1589933838
##                                       obese2Yes
##                                       0.3220127042
##                                    smokerCurrent
##                                       0.6453523887
##                                    smokerFormer
##                                       0.1939432550
##                                       strokeYes
##                                       0.6776804379
##                                             pcPC
##                                       0.1029704091
##          scale(age, scale = FALSE):diabeticYes
##                                      -0.0150328163
##   scale(age, scale = FALSE):educ9th-11th grade
##                                       0.0055464480
##      scale(age, scale = FALSE):educHS graduate
##                                       0.0060640033
##     scale(age, scale = FALSE):educSome college
##                                       0.0006781886
## scale(age, scale = FALSE):educCollege graduate
##                                       0.0175761083
##      scale(age, scale = FALSE):hypertensionYes
##                                      -0.0221211970
##            scale(age, scale = FALSE):strokeYes
##                                      -0.0160291677

sp2 <-
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker + stroke +
      scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke + pc,
    data = mydata[mydata$inmodel4 == 1 &
                    mydata$permth_exm > 0, ],
    k = 2,
    scale = "hazard"
  )

coefficients(sp2)
##                                       gamma0
##                                   -8.5599930650
##                                       gamma1
```

```
##                                     1.3177665707
##                                            gamma2
##                                     0.1535177779
##                                            gamma3
##                                    -0.2746807605
##                       scale(age, scale = FALSE)
##                                     0.1001397819
##                                      diabeticYes
##                                     0.4230474003
##                                  educ9th-11th grade
##                                     0.0269367571
##                                   educHS graduate
##                                    -0.1071644914
##                                  educSome college
##                                    -0.1205292615
##                               educCollege graduate
##                                    -0.5712980383
##                                   hypertensionYes
##                                     0.3771028145
##                                 marital2Separated
##                                     0.3553077949
##                                   marital2Single
##                                     0.5728985915
##                                   underweightYes
##                                     0.7814359563
##                                   overweight2Yes
##                                    -0.1594161802
##                                       obese2Yes
##                                     0.3247326201
##                                   smokerCurrent
##                                     0.6497940626
##                                    smokerFormer
##                                     0.1943978216
##                                       strokeYes
##                                     0.6793412857
##                                             pcPC
##                                     0.1067594941
##          scale(age, scale = FALSE):diabeticYes
##                                    -0.0151029413
##   scale(age, scale = FALSE):educ9th-11th grade
##                                     0.0054665820
##      scale(age, scale = FALSE):educHS graduate
##                                     0.0059171257
##     scale(age, scale = FALSE):educSome college
##                                     0.0006175035
## scale(age, scale = FALSE):educCollege graduate
##                                     0.0172220692
##      scale(age, scale = FALSE):hypertensionYes
##                                    -0.0220700167
##            scale(age, scale = FALSE):strokeYes
##                                    -0.0158829967

sp1_td <-
```

```
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + gamma1(scale(age, scale = FALSE)) +
      diabetic + educ + hypertension + marital2 + underweight + overweight2 + obese2 +
      smoker + stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke + pc,
    data = mydata[mydata$inmodel4 == 1 &
                    mydata$permth_exm > 0, ],
    k = 1,
    scale = "hazard"
  )

coefficients(sp1_td)
##                                       gamma0
##                                -8.0243667838
##                                       gamma1
##                                 1.0788901278
##                                       gamma2
##                                -0.0318099829
##                       scale(age, scale = FALSE)
##                                 0.0860490964
##                                  diabeticYes
##                                 0.4180838641
##                            educ9th-11th grade
##                                 0.0262595286
##                               educHS graduate
##                                -0.1113664799
##                             educSome college
##                                -0.1252651042
##                          educCollege graduate
##                                -0.5756534873
##                               hypertensionYes
##                                 0.3760493283
##                              marital2Separated
##                                 0.3527328214
##                               marital2Single
##                                 0.5660393922
##                               underweightYes
##                                 0.7749386053
##                               overweight2Yes
##                                -0.1593905850
##                                   obese2Yes
##                                 0.3204415714
##                               smokerCurrent
##                                 0.6454888191
##                                smokerFormer
##                                 0.1947923029
##                                   strokeYes
##                                 0.6744686299
##                                       pcPC
##                                 0.1039072933
##          scale(age, scale = FALSE):diabeticYes
##                                -0.0148369718
##   scale(age, scale = FALSE):educ9th-11th grade
```

```
##                                     0.0055445567
##      scale(age, scale = FALSE):educHS graduate
##                                     0.0060529199
##     scale(age, scale = FALSE):educSome college
##                                     0.0006679483
## scale(age, scale = FALSE):educCollege graduate
##                                     0.0174268737
##      scale(age, scale = FALSE):hypertensionYes
##                                    -0.0221823752
##            scale(age, scale = FALSE):strokeYes
##                                    -0.0157141638
##              gamma1(scale(age, scale = FALSE))
##                                     0.0028642796
```

```
sp2_td <-
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + gamma1(scale(age, scale = FALSE)) +
      diabetic + educ + hypertension + marital2 + underweight + overweight2 + obese2 +
      smoker + stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke + pc,
    data = mydata[mydata$inmodel4 == 1 &
                    mydata$permth_exm > 0, ],
    k = 2,
    scale = "hazard"
  )

coefficients(sp2_td)
##                                         gamma0
##                                    -8.3660413590
##                                         gamma1
##                                     1.2799579362
##                                         gamma2
##                                     0.1616504223
##                                         gamma3
##                                    -0.2883824319
##                          scale(age, scale = FALSE)
##                                     0.0823726365
##                                      diabeticYes
##                                     0.4196411478
##                                educ9th-11th grade
##                                     0.0279963849
##                                   educHS graduate
##                                    -0.1062768055
##                                  educSome college
##                                    -0.1200766194
##                               educCollege graduate
##                                    -0.5698396820
##                                  hypertensionYes
##                                     0.3767560412
##                                 marital2Separated
##                                     0.3563582806
##                                   marital2Single
##                                     0.5722696597
```

Table 1: AIC of Parametric Spline Models (Smaller Better)

Method	AIC
1 Knot	23168.45
2 Knots	23162.13
1 Knot, Time Dep.	23169.24
2 Knots, Time Dep.	23162.07

```
##                                     underweightYes
##                                       0.7819615927
##                                     overweight2Yes
##                                      -0.1599757407
##                                         obese2Yes
##                                       0.3239444030
##                                      smokerCurrent
##                                       0.6498860634
##                                      smokerFormer
##                                       0.1954640531
##                                         strokeYes
##                                       0.6750132057
##                                             pcPC
##                                       0.1081253146
##          scale(age, scale = FALSE):diabeticYes
##                                      -0.0148531620
##   scale(age, scale = FALSE):educ9th-11th grade
##                                       0.0054506718
##      scale(age, scale = FALSE):educHS graduate
##                                       0.0058642918
##     scale(age, scale = FALSE):educSome college
##                                       0.0006239544
## scale(age, scale = FALSE):educCollege graduate
##                                       0.0169864940
##      scale(age, scale = FALSE):hypertensionYes
##                                      -0.0221225836
##            scale(age, scale = FALSE):strokeYes
##                                      -0.0154470886
##              gamma1(scale(age, scale = FALSE))
##                                       0.0037257171

performance <-
  data.frame(
    "Method" = c("1 Knot", "2 Knots", "1 Knot, Time Dep.", "2 Knots, Time Dep."),
    "AIC" = c(sp1$AIC, sp2$AIC, sp1_td$AIC, sp2_td$AIC)
  )

kable(performance, caption = "AIC of Parametric Spline Models (Smaller Better)")
```

All of the above models seem pretty comparable in terms of AIC. To balance simplicity and AIC, we decide to go with the two knot, time-independent model (sp2) as our candidate parametric spline model.

Sensitivity Analyses: Prostate Cancer Interactions

To assess the effect of prostate cancer on our treatment predictions, we refit the models above without prostate cancer to see if it made a difference in model quality. In addition, we use the linear predictor from these models in a second model that also includes prostate cancer and an interaction between prostate cancer and the linear predictor. None of these sensitivity analyses indicated significantly different effects for prostate cancer and and non-prostate cancer patients. Nonetheless, we left prostate cancer in the model so that we could adjust for it when making predictions for patients.

```
cox_nopc <-
  coxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker +
      stroke + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
    data = mydata[mydata$inmodel2 == 1, ]
  )

summary(cox_nopc)
## Call:
## coxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     diabetic + educ + hypertension + marital2 + underweight +
##     overweight2 + obese2 + smoker + stroke + scale(age, scale = FALSE) *
##     diabetic + scale(age, scale = FALSE) * educ + scale(age,
##     scale = FALSE) * hypertension + scale(age, scale = FALSE) *
##     stroke, data = mydata[mydata$inmodel2 == 1, ])
##
##   n= 7329, number of events= 1769
##
##                                                       coef  exp(coef)   se(coef)
## scale(age, scale = FALSE)                         0.1011478  1.1064401  0.0058905
## diabeticYes                                       0.4088065  1.5050205  0.0744916
## educ9th-11th grade                                0.0323301  1.0328584  0.1028036
## educHS graduate                                  -0.1043667  0.9008949  0.1002808
## educSome college                                 -0.1184134  0.8883288  0.1000585
## educCollege graduate                             -0.5761102  0.5620805  0.1262042
## hypertensionYes                                   0.3841851  1.4684172  0.0690912
## marital2Separated                                 0.3564381  1.4282331  0.0549903
## marital2Single                                    0.5711461  1.7702949  0.1062769
## underweightYes                                    0.7823277  2.1865560  0.1882949
## overweight2Yes                                   -0.1643605  0.8484361  0.0560672
## obese2Yes                                         0.3132227  1.3678261  0.1363012
## smokerCurrent                                     0.6465125  1.9088719  0.0716246
## smokerFormer                                      0.1943739  1.2145503  0.0561139
## strokeYes                                         0.6825920  1.9790008  0.1218634
## scale(age, scale = FALSE):diabeticYes            -0.0136178  0.9864745  0.0054893
## scale(age, scale = FALSE):educ9th-11th grade      0.0040511  1.0040593  0.0071837
## scale(age, scale = FALSE):educHS graduate         0.0044858  1.0044959  0.0068754
## scale(age, scale = FALSE):educSome college       -0.0005872  0.9994130  0.0069762
## scale(age, scale = FALSE):educCollege graduate    0.0174766  1.0176302  0.0082398
## scale(age, scale = FALSE):hypertensionYes        -0.0221483  0.9780952  0.0047158
## scale(age, scale = FALSE):strokeYes              -0.0157704  0.9843533  0.0081179
##                                                       z Pr(>|z|)
## scale(age, scale = FALSE)                        17.171  < 2e-16 ***
```

```
## diabeticYes                                        5.488 4.07e-08 ***
## educ9th-11th grade                                 0.314 0.753153
## educHS graduate                                   -1.041 0.297994
## educSome college                                  -1.183 0.236634
## educCollege graduate                              -4.565 5.00e-06 ***
## hypertensionYes                                    5.561 2.69e-08 ***
## marital2Separated                                  6.482 9.06e-11 ***
## marital2Single                                     5.374 7.70e-08 ***
## underweightYes                                     4.155 3.26e-05 ***
## overweight2Yes                                    -2.931 0.003373 **
## obese2Yes                                          2.298 0.021561 *
## smokerCurrent                                      9.026  < 2e-16 ***
## smokerFormer                                       3.464 0.000532 ***
## strokeYes                                          5.601 2.13e-08 ***
## scale(age, scale = FALSE):diabeticYes             -2.481 0.013108 *
## scale(age, scale = FALSE):educ9th-11th grade       0.564 0.572799
## scale(age, scale = FALSE):educHS graduate          0.652 0.514116
## scale(age, scale = FALSE):educSome college        -0.084 0.932918
## scale(age, scale = FALSE):educCollege graduate     2.121 0.033922 *
## scale(age, scale = FALSE):hypertensionYes         -4.697 2.65e-06 ***
## scale(age, scale = FALSE):strokeYes               -1.943 0.052056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                   exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                            1.1064     0.9038    1.0937
## diabeticYes                                          1.5050     0.6644    1.3006
## educ9th-11th grade                                   1.0329     0.9682    0.8444
## educHS graduate                                      0.9009     1.1100    0.7401
## educSome college                                     0.8883     1.1257    0.7301
## educCollege graduate                                 0.5621     1.7791    0.4389
## hypertensionYes                                      1.4684     0.6810    1.2824
## marital2Separated                                    1.4282     0.7002    1.2823
## marital2Single                                       1.7703     0.5649    1.4374
## underweightYes                                       2.1866     0.4573    1.5118
## overweight2Yes                                       0.8484     1.1786    0.7601
## obese2Yes                                            1.3678     0.7311    1.0472
## smokerCurrent                                        1.9089     0.5239    1.6589
## smokerFormer                                         1.2146     0.8234    1.0881
## strokeYes                                            1.9790     0.5053    1.5585
## scale(age, scale = FALSE):diabeticYes                0.9865     1.0137    0.9759
## scale(age, scale = FALSE):educ9th-11th grade         1.0041     0.9960    0.9900
## scale(age, scale = FALSE):educHS graduate            1.0045     0.9955    0.9911
## scale(age, scale = FALSE):educSome college           0.9994     1.0006    0.9858
## scale(age, scale = FALSE):educCollege graduate       1.0176     0.9827    1.0013
## scale(age, scale = FALSE):hypertensionYes            0.9781     1.0224    0.9691
## scale(age, scale = FALSE):strokeYes                  0.9844     1.0159    0.9688
##                                                   upper .95
## scale(age, scale = FALSE)                            1.1193
## diabeticYes                                          1.7416
## educ9th-11th grade                                   1.2634
## educHS graduate                                      1.0966
## educSome college                                     1.0808
```

```
## educCollege graduate                                0.7198
## hypertensionYes                                     1.6814
## marital2Separated                                   1.5908
## marital2Single                                      2.1803
## underweightYes                                      3.1626
## overweight2Yes                                      0.9470
## obese2Yes                                           1.7867
## smokerCurrent                                       2.1966
## smokerFormer                                        1.3558
## strokeYes                                           2.5129
## scale(age, scale = FALSE):diabeticYes               0.9971
## scale(age, scale = FALSE):educ9th-11th grade        1.0183
## scale(age, scale = FALSE):educHS graduate           1.0181
## scale(age, scale = FALSE):educSome college          1.0132
## scale(age, scale = FALSE):educCollege graduate      1.0342
## scale(age, scale = FALSE):hypertensionYes           0.9872
## scale(age, scale = FALSE):strokeYes                 1.0001
##
## Concordance= 0.796  (se = 0.005 )
## Likelihood ratio test= 2118  on 22 df,   p=<2e-16
## Wald test            = 1647  on 22 df,   p=<2e-16
## Score (logrank) test = 2179  on 22 df,   p=<2e-16
```

```
rforest_nopc <-
  rfsrc(
    Surv(permth_exm, mortstat) ~ age + arthritis + bronch + diabetic + educ +
      emphysema + hypertension + single + sep + mi_chd + underweight + overweight_ex +
      obese + liver + black + other + smoker + stroke,
    data = mydata_nf[mydata_nf$inmodel3 == 1, ],
    seed = 1259
  )

varlist_nopc <- vimp(rforest_nopc)

print(rforest_nopc)
##                          Sample size: 7268
##                     Number of deaths: 1751
##                      Number of trees: 1000
##            Forest terminal node size: 15
##        Average no. of terminal nodes: 280.466
## No. of variables tried at each split: 5
##               Total no. of variables: 18
##        Resampling used to grow trees: swor
##     Resample size used to grow trees: 4593
##                             Analysis: RSF
##                               Family: surv
##                       Splitting rule: logrank *random*
##        Number of random split points: 10
##                           Error rate: 21.01%
print(varlist_nopc$importance)
##           age     arthritis        bronch      diabetic          educ
##  1.658130e-01  1.263971e-04  1.313128e-04  3.761566e-03  4.938635e-03
##     emphysema  hypertension        single           sep        mi_chd
```

```
##  5.090123e-03  3.751575e-03  1.909588e-03  3.194122e-03  3.428488e-03
##   underweight overweight_ex         obese         liver         black
##  7.980740e-04  1.055408e-03  6.536337e-06 -9.745491e-05  4.467297e-04
##         other        smoker        stroke
##  1.409717e-03  4.430420e-03  5.843787e-03

sp_nopc <-
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + diabetic + educ +
      hypertension + marital2 + underweight + overweight2 + obese2 + smoker + stroke +
      scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * hypertension + scale(age, scale = FALSE) * stroke,
    data = mydata[mydata$inmodel2 == 1 &
                    mydata$permth_exm > 0, ],
    k = 2,
    scale = "hazard"
  )

coefficients(sp_nopc)
##                                      gamma0
##                                -8.539383e+00
##                                      gamma1
##                                 1.311593e+00
##                                      gamma2
##                                 1.532790e-01
##                                      gamma3
##                                -2.749182e-01
##                    scale(age, scale = FALSE)
##                                 1.006057e-01
##                                  diabeticYes
##                                 4.096504e-01
##                           educ9th-11th grade
##                                 3.169934e-02
##                              educHS graduate
##                                -1.050295e-01
##                             educSome college
##                                -1.185256e-01
##                         educCollege graduate
##                                -5.770338e-01
##                              hypertensionYes
##                                 3.828300e-01
##                            marital2Separated
##                                 3.574916e-01
##                              marital2Single
##                                 5.738128e-01
##                              underweightYes
##                                 7.510762e-01
##                              overweight2Yes
##                                -1.625167e-01
##                                  obese2Yes
##                                 3.140476e-01
##                              smokerCurrent
##                                 6.478300e-01
```

```
##                                     smokerFormer
##                                     1.932591e-01
##                                        strokeYes
##                                     6.819018e-01
##           scale(age, scale = FALSE):diabeticYes
##                                    -1.396548e-02
##    scale(age, scale = FALSE):educ9th-11th grade
##                                     4.693936e-03
##       scale(age, scale = FALSE):educHS graduate
##                                     5.005176e-03
##      scale(age, scale = FALSE):educSome college
##                                    -1.007750e-06
## scale(age, scale = FALSE):educCollege graduate
##                                     1.775517e-02
##       scale(age, scale = FALSE):hypertensionYes
##                                    -2.222403e-02
##             scale(age, scale = FALSE):strokeYes
##                                    -1.538601e-02

pred_cox <- predict(cox_nopc, mydata[mydata$inmodel2 == 1, ])
mydata$pred[mydata$inmodel2 == 1] <- pred_cox

pc_cox_check <-
  coxph(Surv(permth_exm, mortstat) ~ pc + pred + pc * pred,
        data = mydata[mydata$inmodel2 == 1, ])

summary(pc_cox_check)
## Call:
## coxph(formula = Surv(permth_exm, mortstat) ~ pc + pred + pc *
##     pred, data = mydata[mydata$inmodel2 == 1, ])
##
##   n= 7329, number of events= 1769
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## pcPC       0.12714   1.13557  0.20254  0.628    0.530
## pred       0.99616   2.70786  0.02537 39.266   <2e-16 ***
## pcPC:pred -0.01100   0.98906  0.12232 -0.090    0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## pcPC          1.1356     0.8806    0.7635     1.689
## pred          2.7079     0.3693    2.5765     2.846
## pcPC:pred     0.9891     1.0111    0.7782     1.257
##
## Concordance= 0.796  (se = 0.005 )
## Likelihood ratio test= 2120  on 3 df,   p=<2e-16
## Wald test            = 1646  on 3 df,   p=<2e-16
## Score (logrank) test = 1955  on 3 df,   p=<2e-16

predf <- model.matrix(sp_nopc) %*% sp_nopc$res[-(1:4), "est"]
mydata$pred_spline[mydata$inmodel2 == 1 &
                     mydata$permth_exm > 0] <- predf[, 1]
```

```
pc_sp_check <-
  flexsurvspline(
    Surv(permth_exm, mortstat) ~ pc + pred_spline + pc * pred_spline,
    data = mydata[mydata$inmodel2 == 1 &
                    mydata$permth_exm > 0, ],
    k = 2,
    scale = "hazard"
  )

coefficients(pc_sp_check)
##           gamma0           gamma1           gamma2           gamma3
##      -8.54590979       1.31167273       0.15432274      -0.27675540
##             pcPC      pred_spline pcPC:pred_spline
##       0.17691686       0.99727592      -0.03682894
```

Age Assessment

Our collaborators had some concerns about whether an age 40+ population was truly representative of the prostate cancer patient population. In the end, this didn't turn out to be an issue, which should make sense from a statistical perspective, but here is the Cox model we built for men ages 55+:

```
mydata$inmodel5 <-
  ifelse(
    mydata$sex == "Male" &
      mydata$age >= 55 &
      (mydata$cancer == 0 |
         mydata$pc == "PC") &
      !is.na(mydata$age) &
      !is.na(mydata$race2) &
      !is.na(mydata$educ) &
      !is.na(mydata$marital2) &
      !is.na(mydata$emphysema) &
      !is.na(mydata$diabetic) &
      !is.na(mydata$stroke) &
      !is.na(mydata$smoker) &
      !is.na(mydata$BMXBMI) &
      !is.na(mydata$mortstat) & !is.na(mydata$pc),
    1,
    0
  )

cox_55 <-
  coxph(
    Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) + race2 + educ + marital2 +
      emphysema + diabetic + stroke + smoker + underweight + overweight2 + obese2 +
      pc + scale(age, scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
      scale(age, scale = FALSE) * marital2 + race2 * educ,
    data = mydata[mydata$inmodel5 == 1, ]
  )

summary(cox_55)
## Call:
```

```
## coxph(formula = Surv(permth_exm, mortstat) ~ scale(age, scale = FALSE) +
##     race2 + educ + marital2 + emphysema + diabetic + stroke +
##     smoker + underweight + overweight2 + obese2 + pc + scale(age,
##     scale = FALSE) * diabetic + scale(age, scale = FALSE) * educ +
##     scale(age, scale = FALSE) * marital2 + race2 * educ, data = mydata[mydata$inmodel5 ==
##     1, ])
##
##   n= 4876, number of events= 1718
##
##                                                      coef exp(coef)  se(coef)
## scale(age, scale = FALSE)                        0.097916  1.102870  0.007431
## race2NHW                                        -0.053500  0.947906  0.134303
## race2Other                                      -0.331716  0.717691  0.129383
## educ9th-11th grade                              -0.091183  0.912851  0.154480
## educHS graduate                                 -0.318729  0.727073  0.174620
## educSome college                                -0.515068  0.597460  0.187385
## educCollege graduate                            -0.207582  0.812547  0.217880
## marital2Separated                                0.380419  1.462898  0.067764
## marital2Single                                   0.517211  1.677344  0.124443
## emphysemaYes                                     0.600176  1.822440  0.088258
## diabeticYes                                      0.419155  1.520676  0.062098
## strokeYes                                        0.403143  1.496521  0.076842
## smokerCurrent                                    0.526745  1.693412  0.077677
## smokerFormer                                     0.107841  1.113871  0.056271
## underweightYes                                   0.725141  2.065022  0.193311
## overweight2Yes                                  -0.099247  0.905519  0.055893
## obese2Yes                                        0.140908  1.151318  0.163632
## pcPC                                             0.026927  1.027292  0.080507
## scale(age, scale = FALSE):diabeticYes           -0.014460  0.985644  0.007138
## scale(age, scale = FALSE):educ9th-11th grade    -0.006368  0.993652  0.009580
## scale(age, scale = FALSE):educHS graduate        0.012284  1.012360  0.009566
## scale(age, scale = FALSE):educSome college       0.005544  1.005560  0.009464
## scale(age, scale = FALSE):educCollege graduate   0.027523  1.027905  0.010644
## scale(age, scale = FALSE):marital2Separated     -0.016041  0.984087  0.006853
## scale(age, scale = FALSE):marital2Single        -0.046759  0.954317  0.014725
## race2NHW:educ9th-11th grade                      0.266775  1.305747  0.188273
## race2Other:educ9th-11th grade                   -0.141076  0.868423  0.223867
## race2NHW:educHS graduate                         0.179822  1.197005  0.200795
## race2Other:educHS graduate                       0.061703  1.063646  0.253962
## race2NHW:educSome college                        0.388078  1.474145  0.216006
## race2Other:educSome college                      0.395801  1.485573  0.258033
## race2NHW:educCollege graduate                   -0.482167  0.617444  0.241657
## race2Other:educCollege graduate                 -0.175229  0.839265  0.288648
##                                                      z Pr(>|z|)
## scale(age, scale = FALSE)                       13.176  < 2e-16 ***
## race2NHW                                        -0.398 0.690368
## race2Other                                      -2.564 0.010353 *
## educ9th-11th grade                              -0.590 0.555019
## educHS graduate                                 -1.825 0.067960 .
## educSome college                                -2.749 0.005983 **
## educCollege graduate                            -0.953 0.340725
## marital2Separated                                5.614 1.98e-08 ***
## marital2Single                                   4.156 3.24e-05 ***
```

```
## emphysemaYes                                        6.800 1.04e-11 ***
## diabeticYes                                         6.750 1.48e-11 ***
## strokeYes                                           5.246 1.55e-07 ***
## smokerCurrent                                       6.781 1.19e-11 ***
## smokerFormer                                        1.916 0.055307 .
## underweightYes                                      3.751 0.000176 ***
## overweight2Yes                                     -1.776 0.075788 .
## obese2Yes                                           0.861 0.389169
## pcPC                                                0.334 0.738030
## scale(age, scale = FALSE):diabeticYes              -2.026 0.042804 *
## scale(age, scale = FALSE):educ9th-11th grade       -0.665 0.506235
## scale(age, scale = FALSE):educHS graduate           1.284 0.199099
## scale(age, scale = FALSE):educSome college          0.586 0.557962
## scale(age, scale = FALSE):educCollege graduate      2.586 0.009715 **
## scale(age, scale = FALSE):marital2Separated        -2.341 0.019243 *
## scale(age, scale = FALSE):marital2Single           -3.175 0.001496 **
## race2NHW:educ9th-11th grade                         1.417 0.156494
## race2Other:educ9th-11th grade                      -0.630 0.528578
## race2NHW:educHS graduate                            0.896 0.370491
## race2Other:educHS graduate                          0.243 0.808036
## race2NHW:educSome college                           1.797 0.072398 .
## race2Other:educSome college                         1.534 0.125050
## race2NHW:educCollege graduate                      -1.995 0.046015 *
## race2Other:educCollege graduate                    -0.607 0.543805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                    exp(coef) exp(-coef) lower .95
## scale(age, scale = FALSE)                             1.1029     0.9067    1.0869
## race2NHW                                              0.9479     1.0550    0.7285
## race2Other                                            0.7177     1.3934    0.5569
## educ9th-11th grade                                    0.9129     1.0955    0.6744
## educHS graduate                                       0.7271     1.3754    0.5163
## educSome college                                      0.5975     1.6738    0.4138
## educCollege graduate                                  0.8125     1.2307    0.5301
## marital2Separated                                     1.4629     0.6836    1.2810
## marital2Single                                        1.6773     0.5962    1.3143
## emphysemaYes                                          1.8224     0.5487    1.5329
## diabeticYes                                           1.5207     0.6576    1.3464
## strokeYes                                             1.4965     0.6682    1.2873
## smokerCurrent                                         1.6934     0.5905    1.4543
## smokerFormer                                          1.1139     0.8978    0.9976
## underweightYes                                        2.0650     0.4843    1.4138
## overweight2Yes                                        0.9055     1.1043    0.8116
## obese2Yes                                             1.1513     0.8686    0.8354
## pcPC                                                  1.0273     0.9734    0.8773
## scale(age, scale = FALSE):diabeticYes                 0.9856     1.0146    0.9720
## scale(age, scale = FALSE):educ9th-11th grade          0.9937     1.0064    0.9752
## scale(age, scale = FALSE):educHS graduate             1.0124     0.9878    0.9936
## scale(age, scale = FALSE):educSome college            1.0056     0.9945    0.9871
## scale(age, scale = FALSE):educCollege graduate        1.0279     0.9729    1.0067
## scale(age, scale = FALSE):marital2Separated           0.9841     1.0162    0.9710
## scale(age, scale = FALSE):marital2Single              0.9543     1.0479    0.9272
```

```
## race2NHW:educ9th-11th grade                          1.3057     0.7658    0.9028
## race2Other:educ9th-11th grade                        0.8684     1.1515    0.5600
## race2NHW:educHS graduate                             1.1970     0.8354    0.8076
## race2Other:educHS graduate                           1.0636     0.9402    0.6466
## race2NHW:educSome college                            1.4741     0.6784    0.9653
## race2Other:educSome college                          1.4856     0.6731    0.8959
## race2NHW:educCollege graduate                        0.6174     1.6196    0.3845
## race2Other:educCollege graduate                      0.8393     1.1915    0.4767
##                                                    upper .95
## scale(age, scale = FALSE)                            1.1191
## race2NHW                                             1.2333
## race2Other                                           0.9248
## educ9th-11th grade                                   1.2356
## educHS graduate                                      1.0238
## educSome college                                     0.8626
## educCollege graduate                                 1.2454
## marital2Separated                                    1.6707
## marital2Single                                       2.1407
## emphysemaYes                                         2.1666
## diabeticYes                                          1.7175
## strokeYes                                            1.7398
## smokerCurrent                                        1.9719
## smokerFormer                                         1.2437
## underweightYes                                       3.0163
## overweight2Yes                                       1.0104
## obese2Yes                                            1.5866
## pcPC                                                 1.2029
## scale(age, scale = FALSE):diabeticYes                0.9995
## scale(age, scale = FALSE):educ9th-11th grade         1.0125
## scale(age, scale = FALSE):educHS graduate            1.0315
## scale(age, scale = FALSE):educSome college           1.0244
## scale(age, scale = FALSE):educCollege graduate       1.0496
## scale(age, scale = FALSE):marital2Separated          0.9974
## scale(age, scale = FALSE):marital2Single             0.9823
## race2NHW:educ9th-11th grade                          1.8885
## race2Other:educ9th-11th grade                        1.3468
## race2NHW:educHS graduate                             1.7742
## race2Other:educHS graduate                           1.7497
## race2NHW:educSome college                            2.2512
## race2Other:educSome college                          2.4634
## race2NHW:educCollege graduate                        0.9915
## race2Other:educCollege graduate                      1.4777
##
## Concordance= 0.753  (se = 0.006 )
## Likelihood ratio test= 1393  on 33 df,   p=<2e-16
## Wald test            = 1245  on 33 df,   p=<2e-16
## Score (logrank) test = 1489  on 33 df,   p=<2e-16
```

Internal Cross Validation Assessment

Before advancing to external validation, we internally assessed our models' predictive performance using cross validation.

```
times <- seq(6, 168, by = 6)

mydata$age_ctr_40[mydata$inmodel4 == 1] <-
  scale(mydata$age[mydata$inmodel4 == 1], scale = FALSE)
mydata$age_ctr_55[mydata$inmodel5 == 1] <-
  scale(mydata$age[mydata$inmodel5 == 1], scale = FALSE)

#The age 55+ model:
cox_55 <-
  coxph(
    Surv(permth_exm, mortstat) ~ age_ctr_55 + race2 + educ + marital2 + emphysema +
      diabetic + stroke + smoker + underweight + overweight2 + obese2 + pc +
      age_ctr_55 * diabetic + age_ctr_55 * educ + age_ctr_55 * marital2 + race2 * educ,
    data = mydata[mydata$inmodel5 == 1, ],
    x = TRUE,
    y = TRUE
  )

save(list = c("cox_55"), file = "cox_55.RData")

cox55_cind <-
  pec::cindex(
    object = cox_55,
    Surv(permth_exm, mortstat) ~ age_ctr_55 + race2 + educ + marital2 + emphysema +
      diabetic + stroke + smoker + underweight + overweight2 + obese2 + pc +
      age_ctr_55 * diabetic + age_ctr_55 * educ + age_ctr_55 * marital2 + race2 * educ,
    data = mydata[mydata$inmodel5 == 1, ],
    eval.times = times,
    splitMethod = "BootCV"
  )

cox55c_data <-
  data.frame("Years" = c(5, 10, 14), "C" = round(
    c(
      cox55_cind$BootCvCindex$coxph[10],
      cox55_cind$BootCvCindex$coxph[20],
      cox55_cind$BootCvCindex$coxph[28]
    ),
    digits = 3
  ))

kable(cox55c_data)
```

Years	C
5	0.757
10	0.740
14	0.737

```
#The age 40+ model:
cox_40 <-
  coxph(
    Surv(permth_exm, mortstat) ~ age_ctr_40 + diabetic + educ + hypertension +
      marital2 + underweight + overweight2 + obese2 + smoker + stroke +
      age_ctr_40 * diabetic + age_ctr_40 * educ + age_ctr_40 * hypertension +
```

```
      age_ctr_40 * stroke + pc,
    data = mydata[mydata$inmodel4 == 1, ],
    x = TRUE,
    y = TRUE
  )

save(list = c("cox_40"), file = "cox_40.RData")

cox40_cind <-
  pec::cindex(
    object = cox_40,
    formula = Surv(permth_exm, mortstat) ~ age_ctr_40 + diabetic + educ + hypertension +
      marital2 + underweight + overweight2 + obese2 + smoker + stroke + age_ctr_40 *
      diabetic + age_ctr_40 * educ + age_ctr_40 * hypertension + age_ctr_40 *
      stroke + pc,
    data = mydata[mydata$inmodel4 == 1, ],
    eval.times = times,
    splitMethod = "BootCV"
  )

cox40c_data <-
  data.frame("Years" = c(5, 10, 14), "C" = round(
    c(
      cox40_cind$BootCvCindex$coxph[10],
      cox40_cind$BootCvCindex$coxph[20],
      cox40_cind$BootCvCindex$coxph[28]
    ),
    digits = 3
  ))

kable(cox40c_data)
```

Years	C
5	0.796
10	0.792
14	0.787

```
#The random forest:
rforest_40 <-
  rfsrc(
    Surv(permth_exm, mortstat) ~ age + arthritis + bronch + diabetic + educ +
      emphysema + hypertension + single + sep + mi_chd + underweight + overweight_ex +
      obese + liver + black + other + smoker + stroke + pc,
    data = mydata_nf[mydata_nf$inmodel3 == 1, ],
    forest = TRUE
  )

save(list=c("rforest_40"), file="rforest_40.RData")

#These lines are commented out because they take a very long time to run:

#forest_cind <-
 # pec::cindex(
  #  object = rforest_40,
```

```
   # formula = Surv(permth_exm, mortstat) ~ age + arthritis + bronch + diabetic +
    #  educ + emphysema + hypertension + single + sep + mi_chd + underweight +
     # overweight_ex + obese + liver + black + other + smoker + stroke + pc,
    #data = mydata_nf[mydata_nf$inmodel3 == 1, ],
    #eval.times = times,
    #splitMethod = "BootCV"
 # )

#forestc_data <-
#  data.frame("Years" = c(5, 10, 14), "C" = round(
 #   c(
  #    forest_cind$BootCvCindex$rfsrc[10],
   #   forest_cind$BootCvCindex$rfsrc[20],
    #  forest_cind$BootCvCindex$rfsrc[28]
    #),
    #digits = 3
  #))

#kable(forestc_data)

#nrep <- 5
#n_folds <- 5

#testdat <- filter(mydata, inmodel4 == 1, permth_exm > 0)

#source("spline_cindex.R")
#spline_cind <-
 # splinecindex(
  #  data = testdat,
   # formula = as.formula(
    #  "Surv(permth_exm, mortstat) ~ age_ctr_40 + diabetic + educ + hypertension +
     # marital2 + underweight + overweight2 + obese2 + smoker + stroke +
    #  age_ctr_40*diabetic + age_ctr_40*educ + age_ctr_40*hypertension +
    #  age_ctr_40*stroke + pc"
    #),
    #nrep,
    #n_folds,
    #times,
    #k = 2
  #)

#splinec_data <-
 # data.frame("Years" = c(5, 10, 14), "C" = round(c(spline_cind[10], spline_cind[20],
                                                  #  spline_cind[28]), digits = 3))

#kable(splinec_data)

save(list = c("mydata", "mydata_nf"), file = "/Users/ecchase/Desktop/Research/Matt Prostate OCM/PCOtherC

#Now that model building is complete, proceed to OCM_modelvalidation to see how
#prediction models were validated.
```