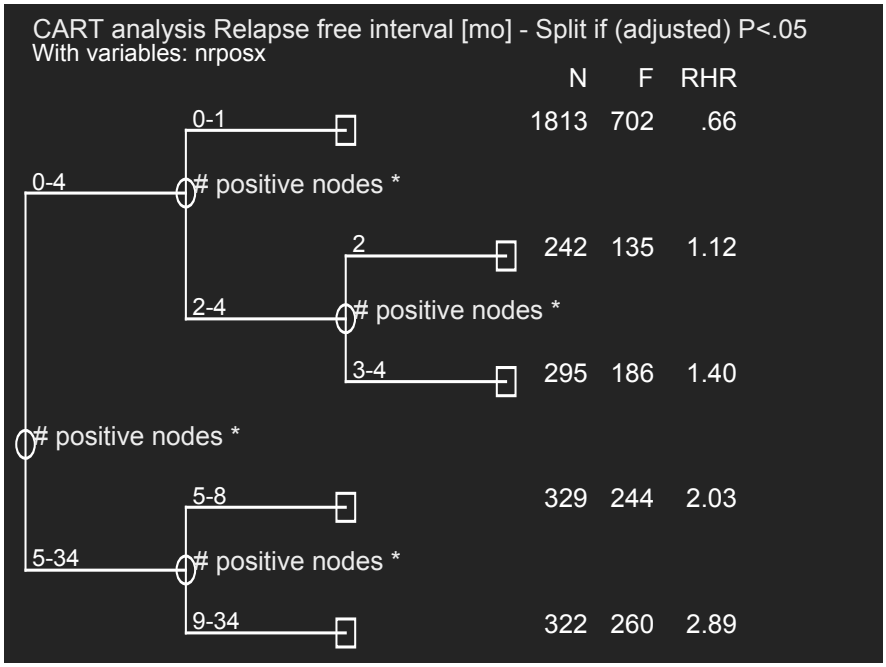
Classification And Regression Tree analysis with Stata

Wim van Putten

University Hospital Rotterdam
Erasmus Medical Center
Daniel den Hoed Cancer Center
Department of Statistics

NL Stata Users meeting, Maastricht, May 23, 2002



Rotterdam Breast Cancer series 3001

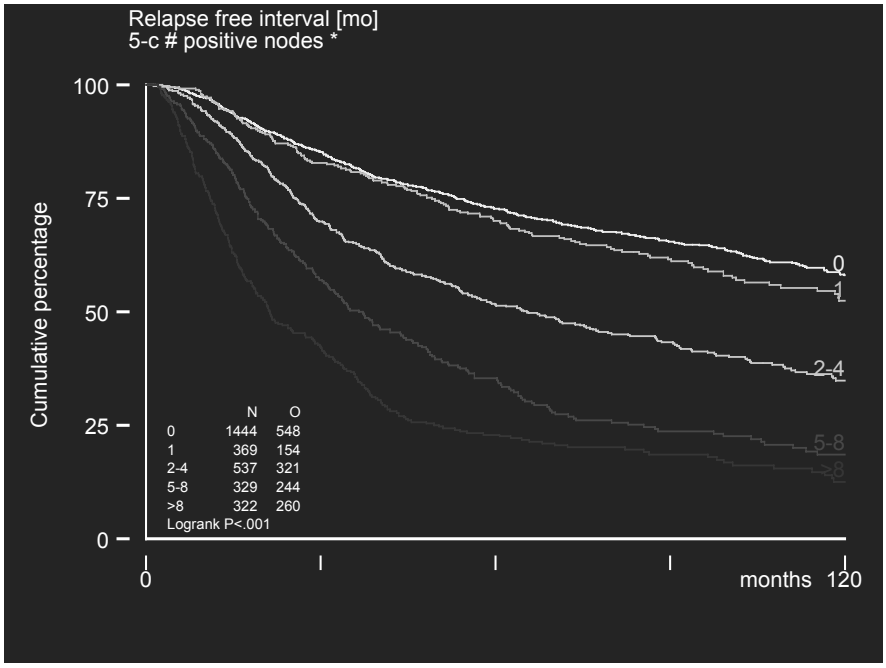
	Adjuvant treatment		
	no	yes	Total
Number	2094	907	3001
Year of surgery			
Range	1978-1993	1979-1993	1978-1993
Median	1988	1989	1988
Age			
Mean;SD	57;13	51;12	55;13
Range	25-90	24-88	24-90
Median	57	50	54

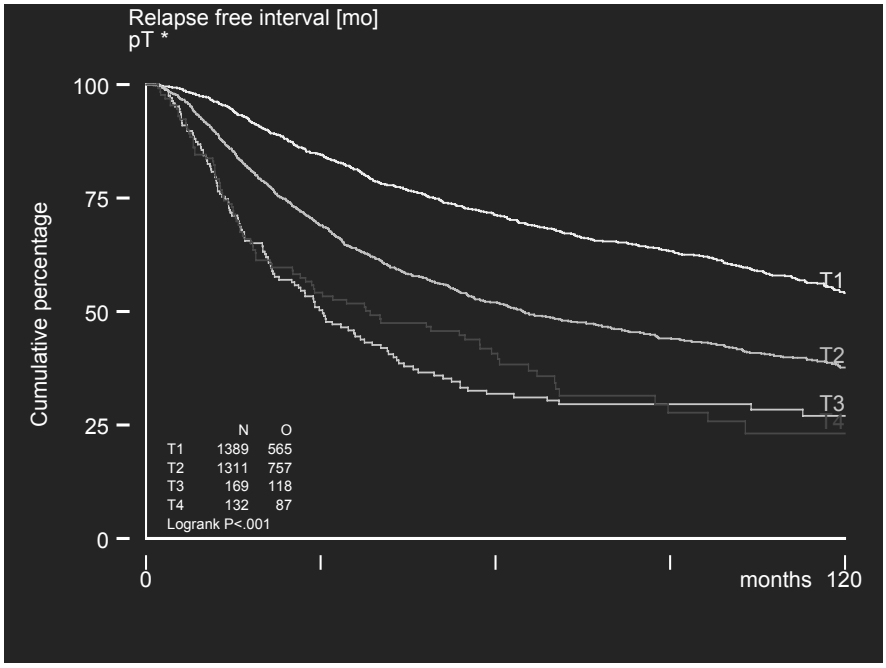
Rotterdam Breast Cancer series 3001

	Adjuvant treatment		
	no	yes	Total
Age	row %		
<=40	58	42	415
>40-<=55	63	37	1165
>55-<=70	77	23	1005
>70	81	19	416
pT *			
T1	76	24	1389
T2	67	33	1311
T3	50	50	169
T4	58	42	132
# positive nodes *			
0	99	1	1444
1	33	67	369
2-4	39	61	537
5-8	53	47	329
>8	48	52	322

```
. sby agek tx np5 adjuvant  ,time(rf) dead(rfi) hea
st(n f at) actlab(RF) flab(REL) at(60, 120)

Relapse free interval [mo]
--------------------------
Events:  Relapse indicator

Time [mnth]                 _____60__ ____120__
                  N  REL  RF%  #at  RF%  #at
age in classes
 <=40           415  256   50  192   33   57
 >40-<=55      1165  567   63  672   48  228
 >55-<=70      1005  516   60  538   44  160
 >70            416  188   59  180   47   38

pT *
 T1            1389  565   71  894   54  273
 T2            1311  757   52  608   38  187
 T3             169  118   32   44   27   17
 T4             132   87   41   36   23    6
```

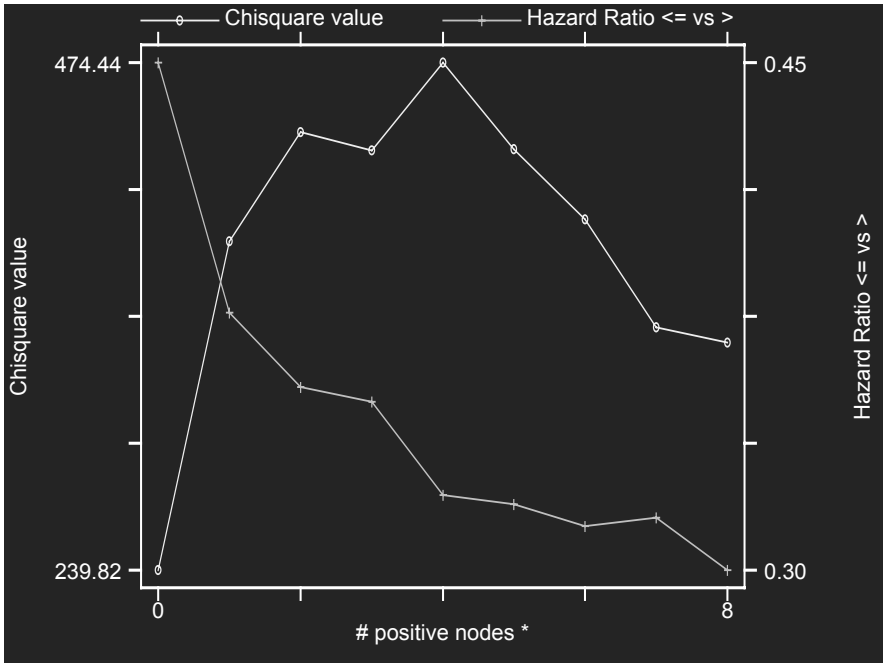
```
Relapse free interval [mo]
--------------------------
Events:  Relapse indicator

Time [mnth]                 _____60__ ____120__
                  N  REL  RF%  #at  RF%  #at

# positive nodes *
 0              1444  548   73  948   58  307
 1               369  154   70  230   52   69
 2-4             537  321   52  250   35   70
 5-8             329  244   35   93   19   20
 >8              322  260   23   61   13   17

Adjuvant treatment
 no             2094 1015   62 1161   47  378
 yes             907  512   53  421   37  105
```


Relapse free interval [mo]
5-c # positive nodes *
Cumulative percentage
100
75
50
25
0
0
months 120
0
1
2-4
5-8
>8
N O
0 1444 548
1 369 154
2-4 537 321
5-8 329 244
>8 322 260
Logrank P<.001


Relapse free interval [mo]
pT *
Cumulative percentage
100
75
50
25
0
0
months 120
T1
T2
T3
T4
N O
T1 1389 565
T2 1311 757
T3 169 118
T4 132 87
Logrank P<.001

CART steps

- Start with full group
- Split (graft) group if splittable
- Repeat until all groups unsplittable
- No pruning or amalgamation of small groups
- Show the results with a tree (+ tables)

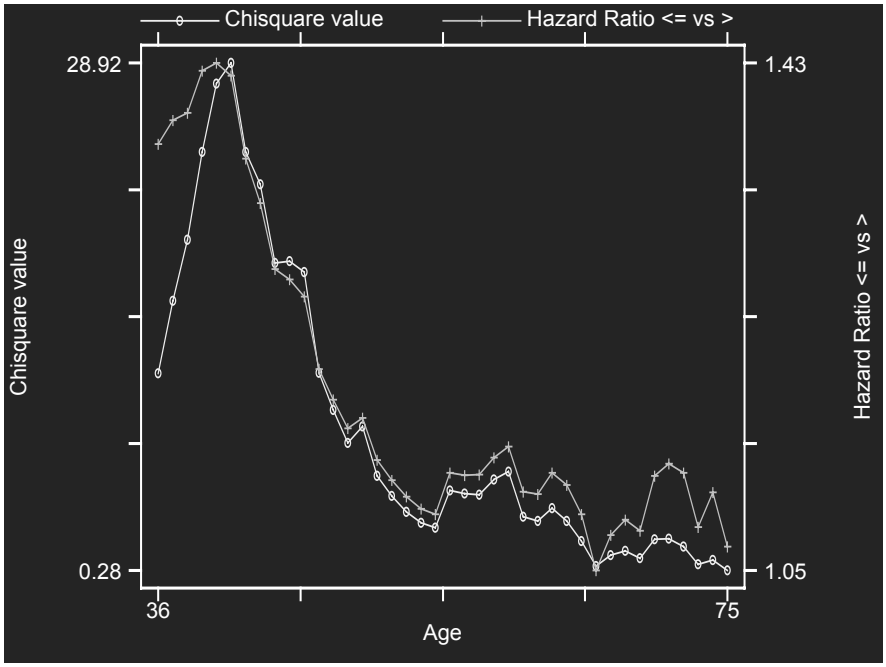


Isotonic regression analysis: # positive nodes

```
. srd nrposx ,time(rf) fail(rfi)

IRA assuming Hazard Rate (O/E)
increasing with nrposx
  nrposx     N     O   HR
--------------------------
  0 -  0  1443   548  0.6
  1 -  1   369   154  0.7
  2 -  2   242   135  1.1
  3 -  4   295   186  1.4
  5 -  5   109    84  1.9
  6 -  6    93    67  2.0
  7 -  8   127    93  2.2
  9 - 10   117    92  2.8
 11 - 16   154   122  2.9
 17 - 30    49    44  3.1
 34 - 34     2     2  8.1
--------------------------
```

Cutpoint for # positive nodes with max Chi square

```
Optimal cutpoint of nrposx at 4
 with chisquare value         474.438
 P-value
  nominal                       3.5e-105
  adjusted (Miller&Siegmund)    4.9e-102
  adjusted (Worsley)           -4.4e-103 ???
  # cutpoints                      9
```



Isotonic regression analysis: Age adjusted for # positive nodes

```
. xxi:srd age ,time(rf) fail(rfi) adjust(i.np5)

IRA assuming Hazard Rate (O/E) decreasing with age
    age       N     O    HR
---------------------------
 24 - 24      1     1  3.90
 25 - 26      4     3  2.95
 27 - 34    133    84  1.48
 35 - 40    277   168  1.29
 41 - 41     47    27  1.21
 42 - 46    417   208  1.00
 47 - 60   1041   516  0.95
 61 - 81   1053   514  0.93
 82 - 82      3     1  0.68
 83 - 83      9     3  0.52
 84 - 90     15     2  0.34
---------------------------
```

Cutpoint for age with max Chi square

```
Optimal cutpoint of age at 41
 with chisquare value          28.921
 P-value
  nominal                       7.5e-08
  adjusted (Miller&Siegmund)    6.6e-06
  adjusted (Worsley)            2.1e-06
  # cutpoints                    40
```

Split criteria

- A group is splittable if
- # failures in group >=MinFail (10)
- and there is a covariate X with cutpoint C :
  - # {obs X<=C} >=MinSize (10)
  - # {obs X>C } >=MinSize (10)
  - logrank test P- value (adjusted) <Pcrit (0.05)
    - based on martingale residuals O-E with optional adjustment for and stratification by other covariates
- Choose X and C with smallest (adj) P-value

References

- Breiman et al, Classification and Regression trees, 1984
- Lausen et al, Informatik, Biometrie und Epidemiologie in Medizin und Biologie 28, 1-13, 1997
- Lausen et al, in Computational Statistics, 483-496, 1994
- Schmoor et al, Stat in Med, 2351-2366, 1993
- Ciampi et al, J Clin Epidemiol, 675-689, 1995
- Ciampi et al, J Clin Epidemiol, 737-748, 1988
- Ciampi et al, Comp Statitics and Data analysis, 185-204, 1986
- LeBlanc et al, Biometrics, 411-425, 1992
- Marubini et al, Stat in Med, 295-303, 1983
- Miller and Siegmund, Biometrics, 1011-1016, 1982
- Worsley, Technometrics, 25, 35-42

Adjustment for multiple testing

- Adjustment for search for optimal cutpoint (Miller & Siegmund, Worsley; see Lausen)
- No adjustment for
  - multiple covariates
  - multiple subgroups

cart.ado

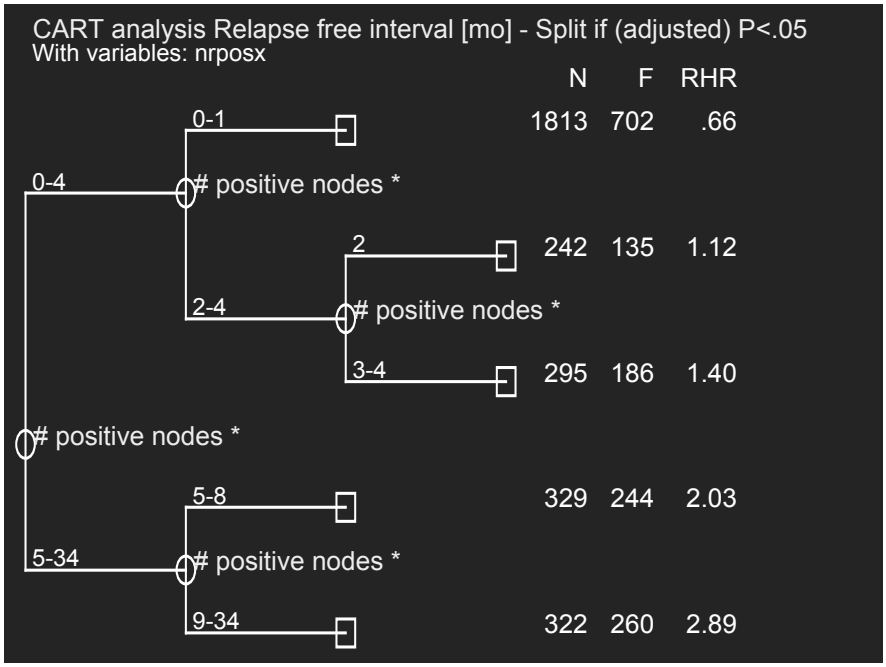
syntax

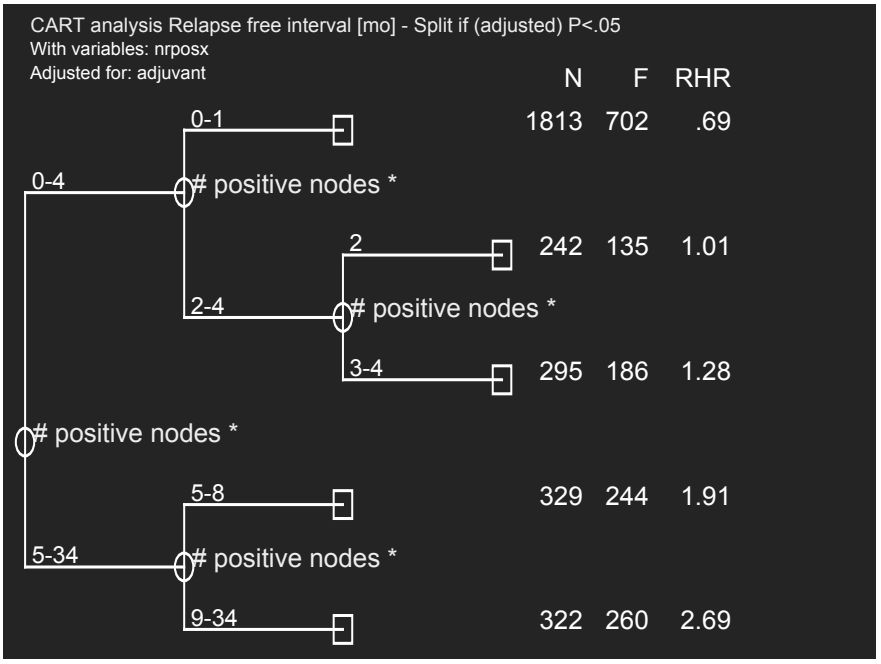
cart varlist [if] [in] , time(var) fail(var)

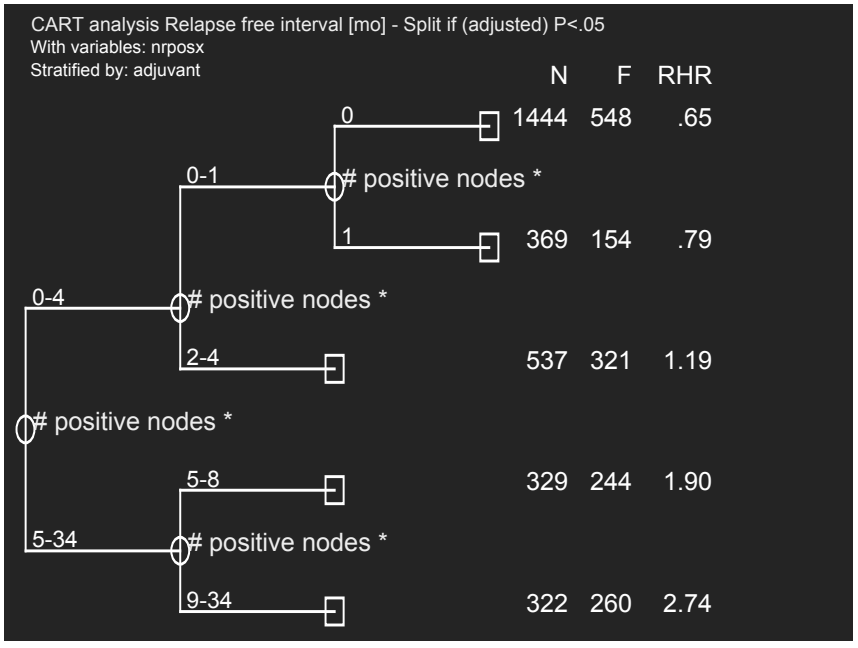
[strata(varlist) adjust(varlist)
pval(real 0.05) pnominal
minsize(int 10) minfail(int 10)

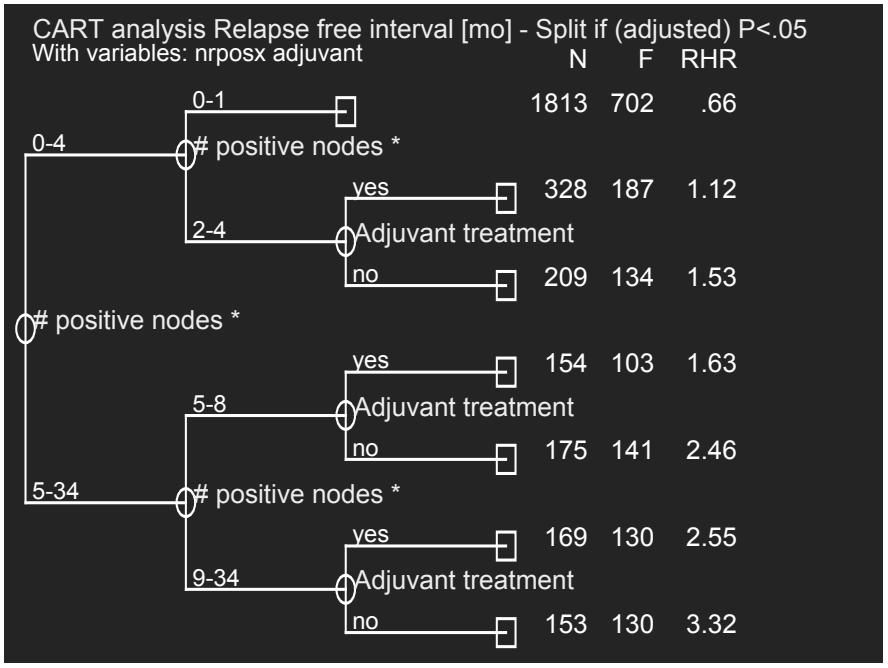
sumby(varlist)
tabby(varlist)
at(string)
name(string)]

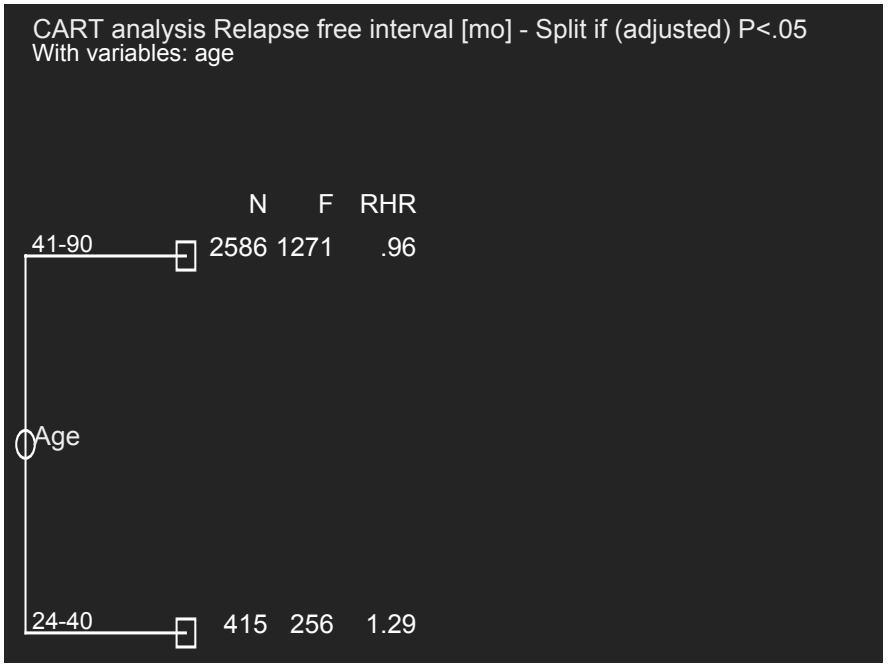
In the examples minsize(50) has been used

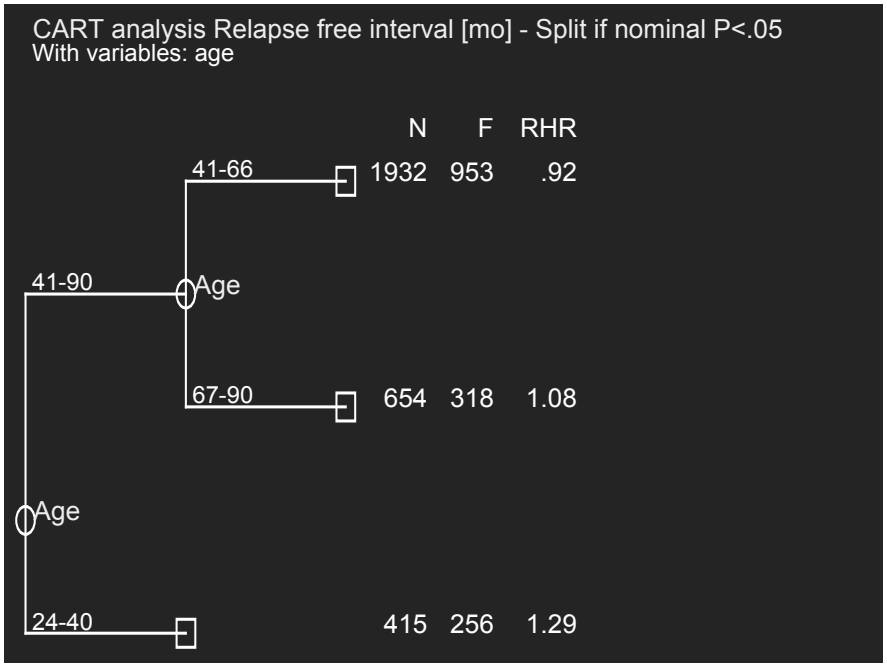


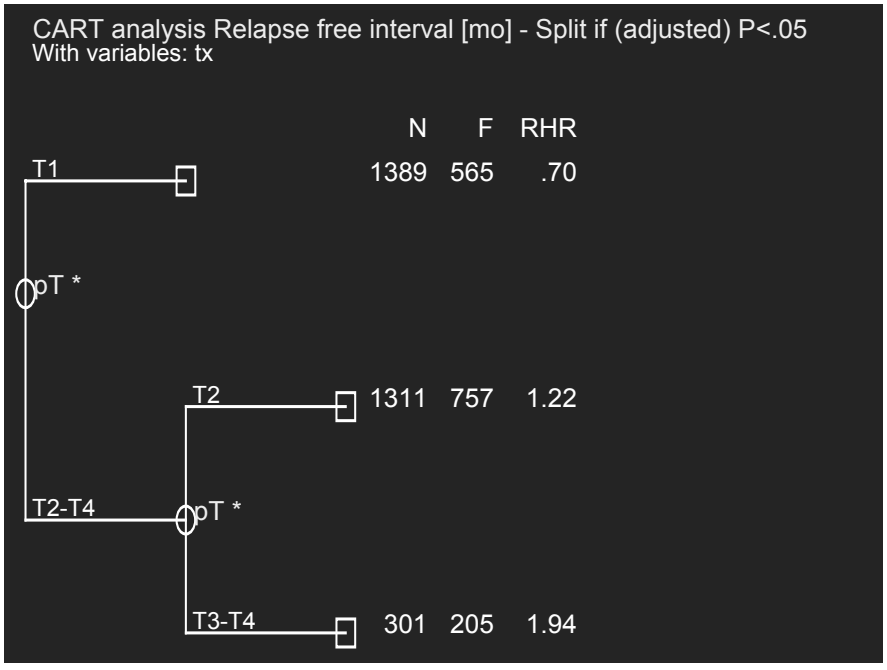

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: nrposx
Adjusted for: adjuvant
N F RHR
0-1
1813 702 .69
0-4
positive nodes *
2
242 135 1.01
2-4
positive nodes *
3-4
295 186 1.28
positive nodes *
5-8
329 244 1.91
5-34
positive nodes *
9-34
322 260 2.69

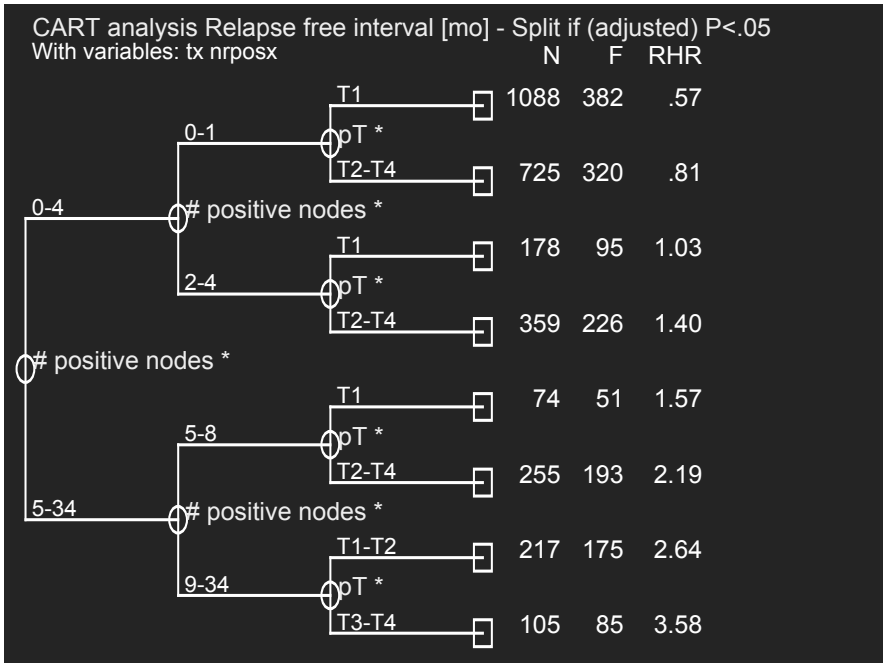

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: nrposx
Stratified by: adjuvant
N F RHR
0
1444 548 .65
0-1
positive nodes *
1
369 154 .79
0-4
positive nodes *
2-4
537 321 1.19
positive nodes *
5-8
329 244 1.90
5-34
positive nodes *
9-34
322 260 2.74

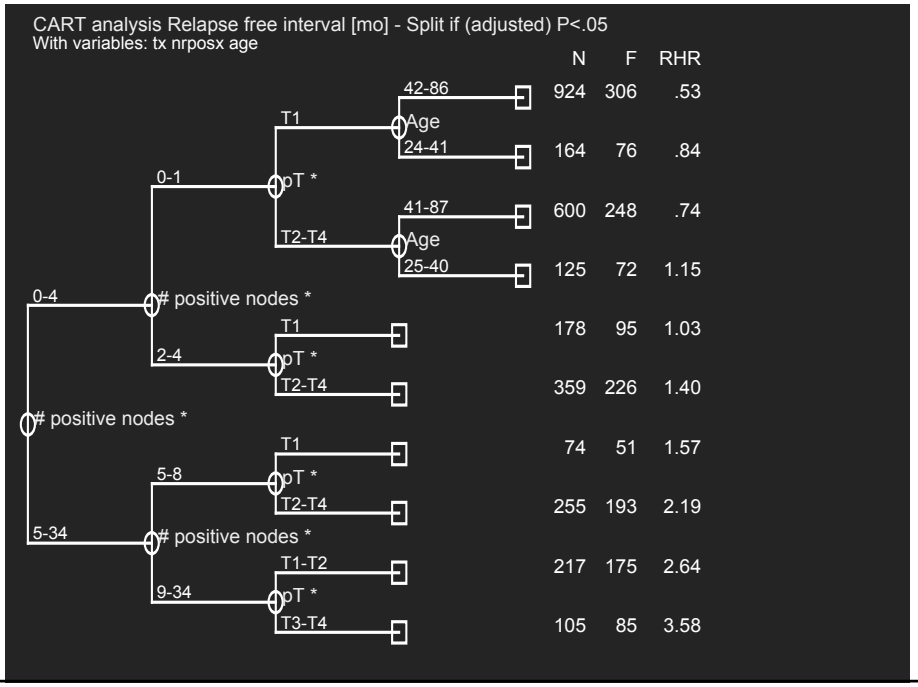

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: nrposx adjuvant
N F RHR
0-1
1813 702 .66
0-4
positive nodes *
yes
328 187 1.12
2-4
Adjuvant treatment
no
209 134 1.53
positive nodes *
yes
154 103 1.63
5-8
Adjuvant treatment
no
175 141 2.46
5-34
positive nodes *
yes
169 130 2.55
9-34
Adjuvant treatment
no
153 130 3.32

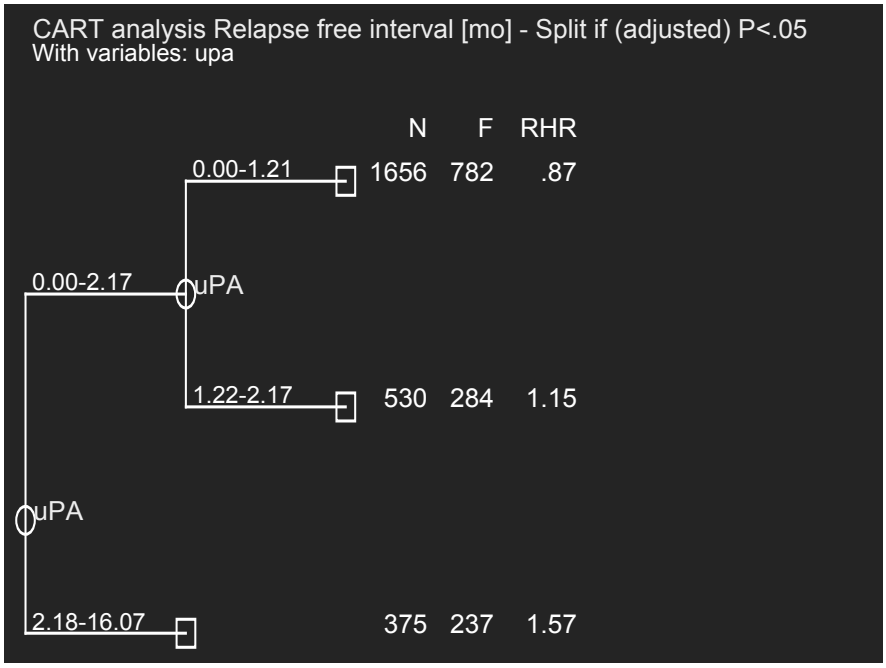

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: age
N F RHR
41-90
2586 1271 .96
Age
24-40
415 256 1.29

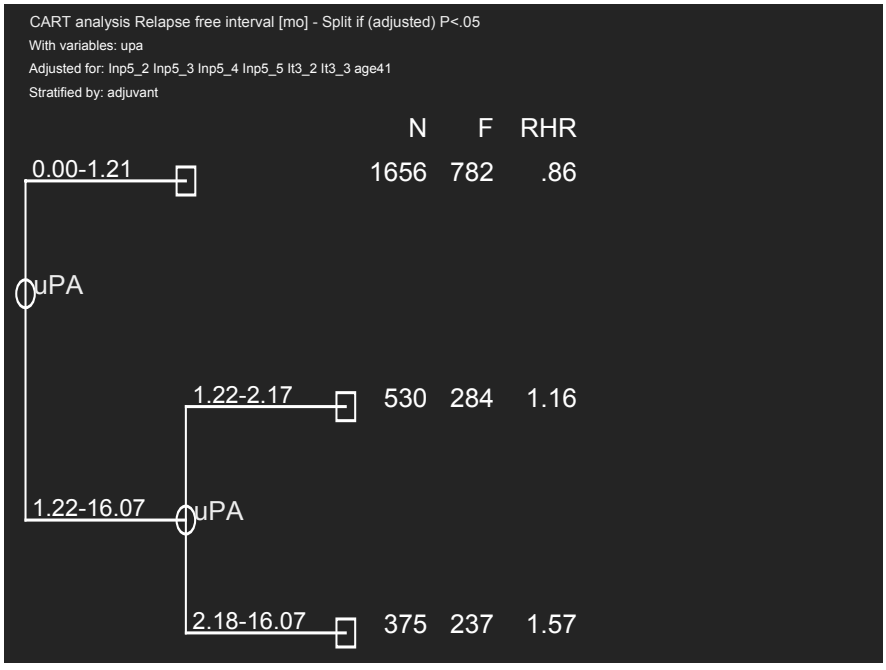

CART analysis Relapse free interval [mo] - Split if nominal P<.05
With variables: age
N F RHR
41-66
1932 953 .92
41-90
Age
67-90
654 318 1.08
Age
24-40
415 256 1.29

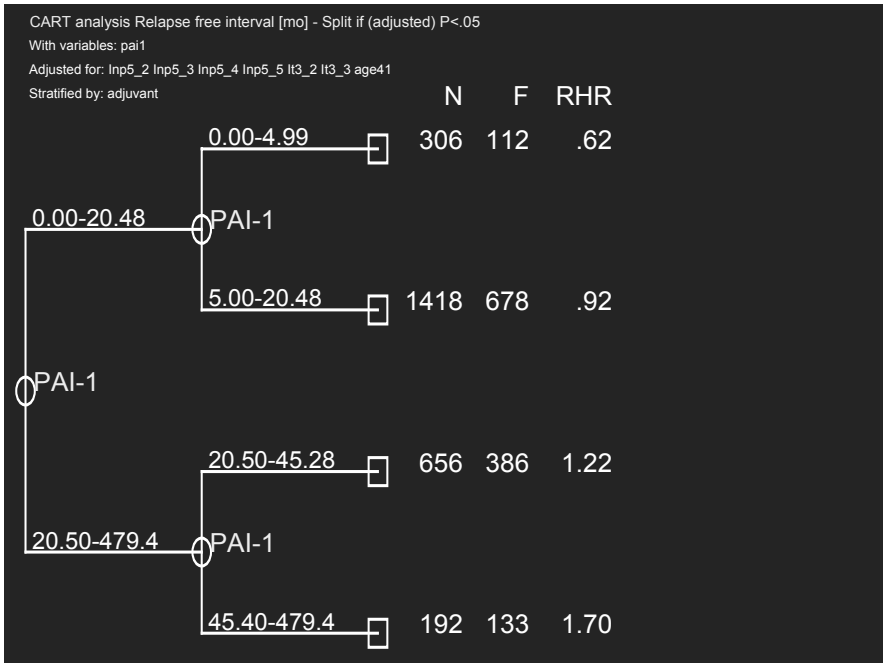

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: tx
N F RHR
T1
1389 565 .70
pT *
T2
1311 757 1.22
T2-T4
pT *
T3-T4
301 205 1.94

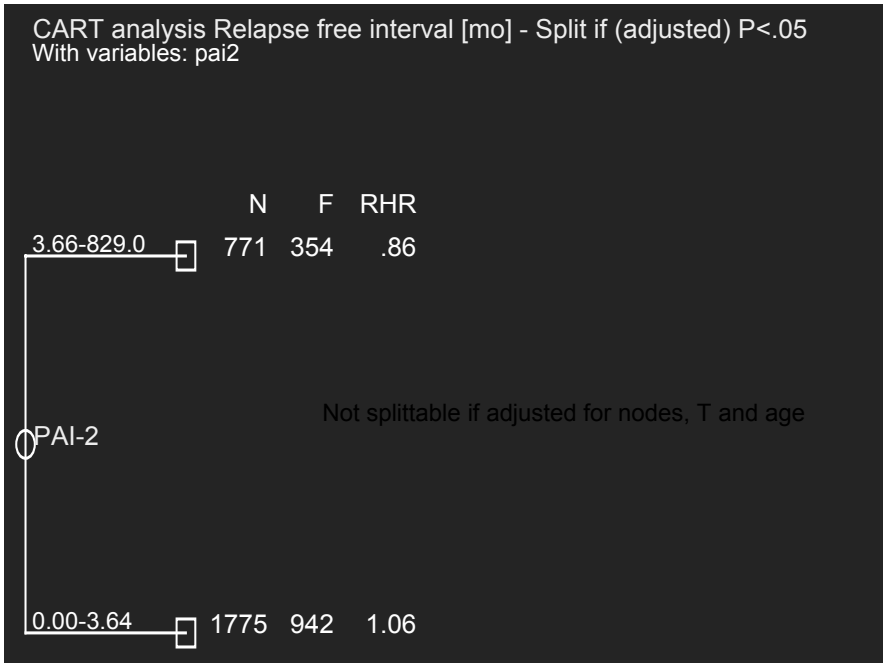

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: tx nrposx
N F RHR
0-4
positive nodes *
0-1
pT *
T1
T2-T4
2-4
pT *
T1
T2-T4
5-34
positive nodes *
5-8
pT *
T1
T2-T4
9-34
pT *
T1-T2
T3-T4
1088 382 .57
725 320 .81
178 95 1.03
359 226 1.40
74 51 1.57
255 193 2.19
217 175 2.64
105 85 3.58

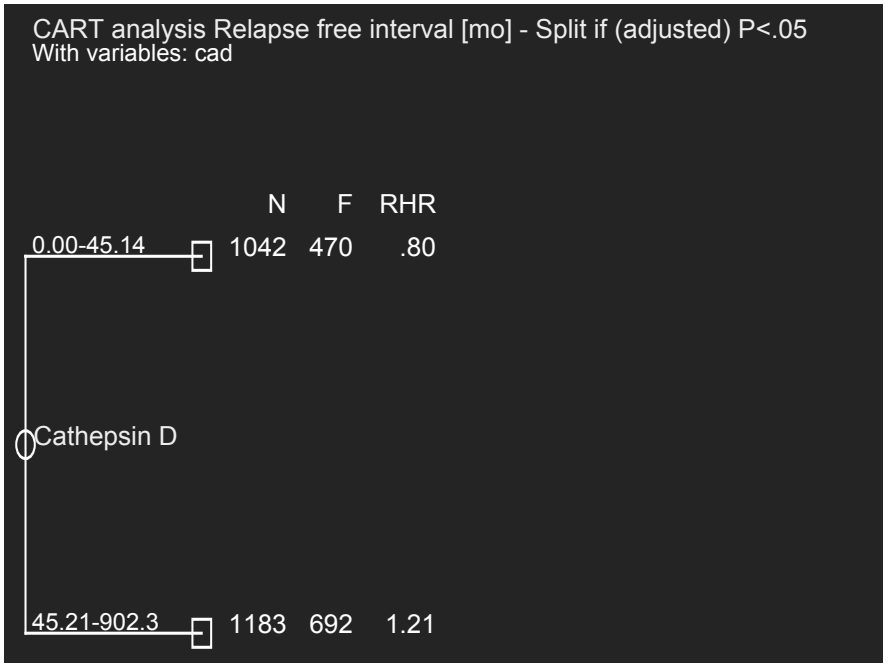

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: tx nrposx age
N F RHR
0-4
positive nodes *
0-1
pT *
T1
Age
42-86
24-41
T2-T4
Age
41-87
25-40
2-4
pT *
T1
T2-T4
5-34
positive nodes *
5-8
pT *
T1
T2-T4
9-34
pT *
T1-T2
T3-T4
924 306 .53
164 76 .84
600 248 .74
125 72 1.15
178 95 1.03
359 226 1.40
74 51 1.57
255 193 2.19
217 175 2.64
105 85 3.58

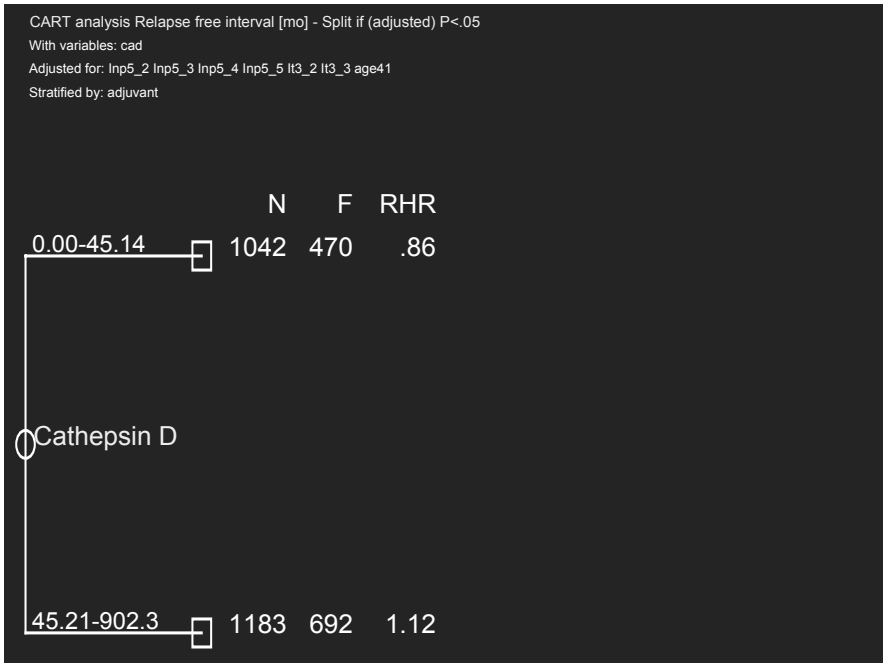

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: upa
N F RHR
0.00-1.21
1656 782 .87
0.00-2.17
uPA
1.22-2.17
530 284 1.15
uPA
2.18-16.07
375 237 1.57

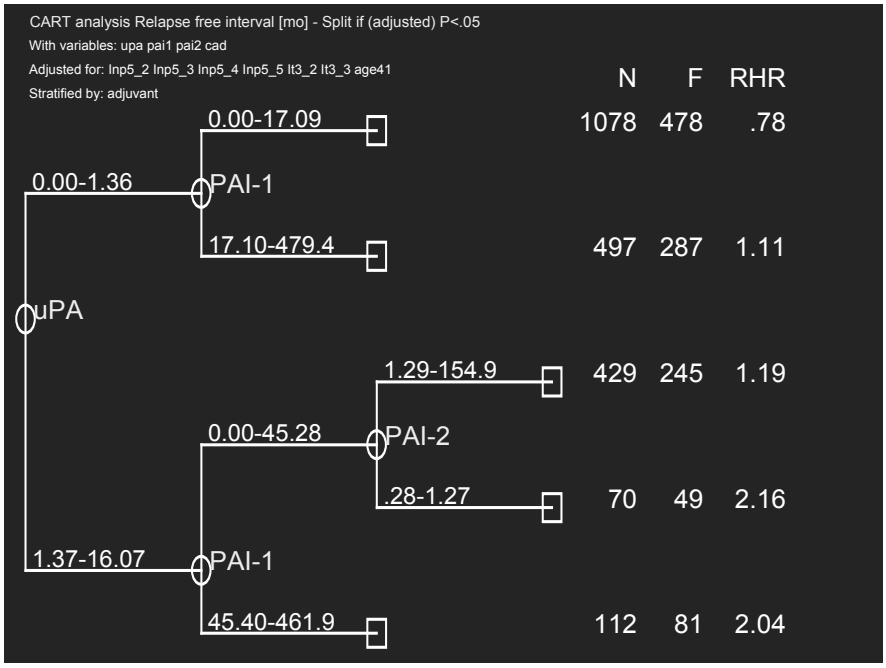

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: upa
Adjusted for: lnp5_2 lnp5_3 lnp5_4 lnp5_5 lt3_2 lt3_3 age41
Stratified by: adjuvant
N F RHR
0.00-1.21
1656 782 .86
uPA
1.22-2.17
530 284 1.16
1.22-16.07
uPA
2.18-16.07
375 237 1.57

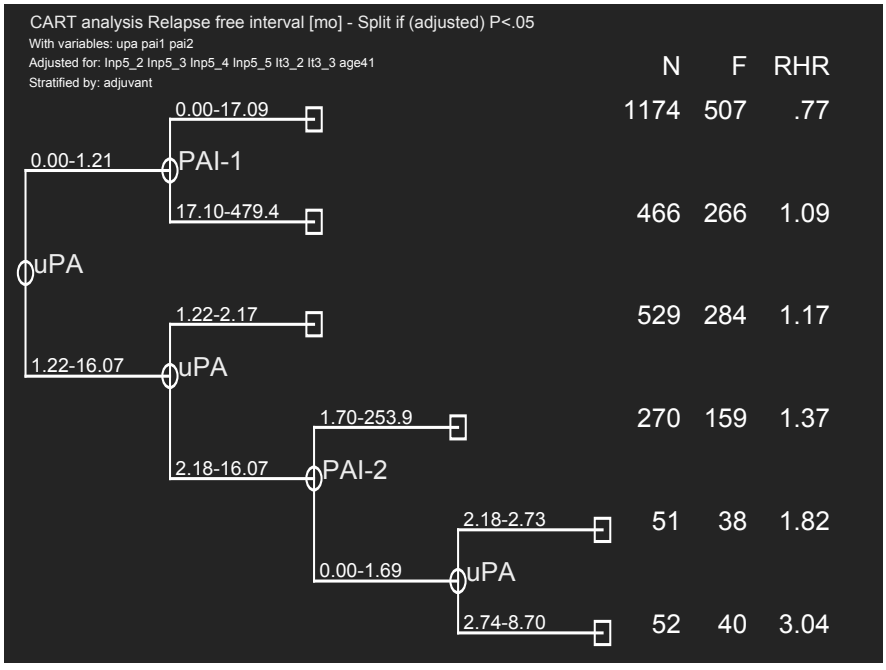

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: pai1
Adjusted for: Inp5_2 Inp5_3 Inp5_4 Inp5_5 It3_2 It3_3 age41
Stratified by: adjuvant
N F RHR
0.00-4.99
306 112 .62
0.00-20.48
PAI-1
5.00-20.48
1418 678 .92
PAI-1
20.50-45.28
656 386 1.22
20.50-479.4
PAI-1
45.40-479.4
192 133 1.70

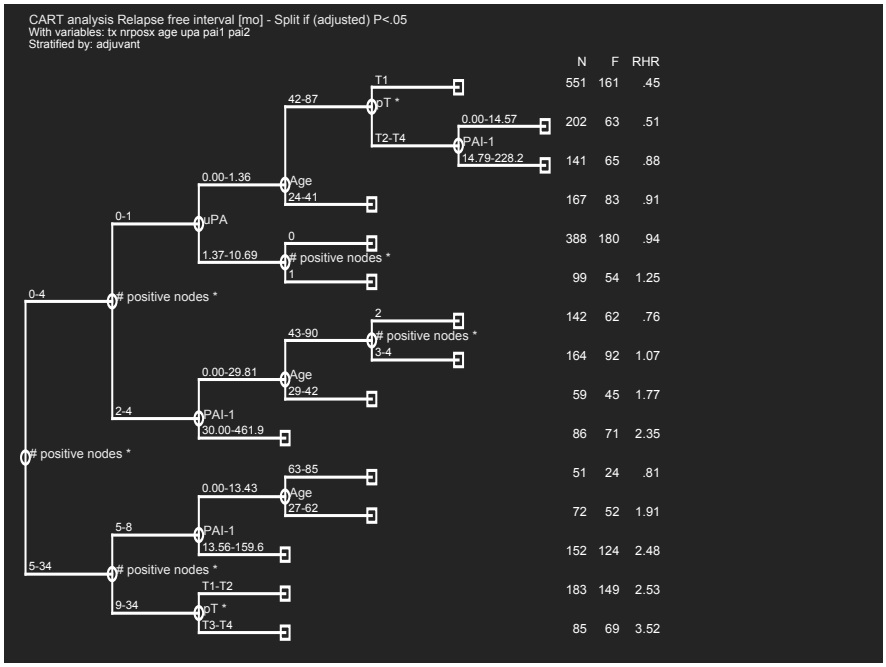

CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: pai2
N F RHR
3.66-829.0
771 354 .86
Not splittable if adjusted for nodes, T and age
PAI-2
0.00-3.64
1775 942 1.06


CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: cad
N F RHR
0.00-45.14
1042 470 .80
Cathepsin D
45.21-902.3
1183 692 1.21


CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: cad
Adjusted for: lnp5_2 lnp5_3 lnp5_4 lnp5_5 lt3_2 lt3_3 age41
Stratified by: adjuvant
N F RHR
0.00-45.14
1042 470 .86
Cathepsin D
45.21-902.3
1183 692 1.12


CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: upa pai1 pai2 cad
Adjusted for: lnp5_2 lnp5_3 lnp5_4 lnp5_5 lt3_2 lt3_3 age41
Stratified by: adjuvant
N F RHR
0.00-17.09
1078 478 .78
0.00-1.36
PAI-1
17.10-479.4
497 287 1.11
uPA
1.29-154.9
429 245 1.19
0.00-45.28
PAI-2
.28-1.27
70 49 2.16
1.37-16.07
PAI-1
45.40-461.9
112 81 2.04


CART analysis Relapse free interval [mo] - Split if (adjusted) P<.05
With variables: upa pai1 pai2
Adjusted for: lnp5_2 lnp5_3 lnp5_4 lnp5_5 lt3_2 lt3_3 age41
Stratified by: adjuvant
N F RHR
0.00-17.09
1174 507 .77
0.00-1.21
PAI-1
17.10-479.4
466 266 1.09
uPA
1.22-2.17
529 284 1.17
1.22-16.07
uPA
1.70-253.9
270 159 1.37
2.18-16.07
PAI-2
2.18-2.73
51 38 1.82
0.00-1.69
uPA
2.74-8.70
52 40 3.04



Conclusions

- CART is a useful exploratory tool
- Helpful in case of one or a few covariates
- A way to choose cutpoints (if you want to)
- Model is not parsimonious
- Overfits “interactions”
- Underfits small/moderate general effects
- Probably best be used in addition to a regression model

CART can be done in Stata

(next week)