Exploratory subgroup identification in the heterogeneous Cox model: A relatively simple procedure

Larry F. León , Thomas Jemielita,
Zifang Guo, Rachel Marceau West, and Keaven Anderson
Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, USA

SUPPLEMENTARY MATERIALS

Section S1.1 considers the accuracy of the $N(\beta, 8/d)$ asymptotic approximation to the random splits of the FS algorithm where the sample size is $N = 60$. Section S2.1 provides an additional analysis of the GBSG dataset. Section S2.2 provides an additional analysis of the ACTG-175 dataset. Sections S2.3 and S2.4 provide additional analyses of the ACTG-175 dataset where we artificially add twenty $N(0,1)$ baseline factors as candidates: In Section S2.3 the FS algorithm does not include lasso, whereas in Section S2.4 lasso is included. Section S2.5 evaluates the use of aspirin in the systolic heart failure data (Hsich et al., 2011) available in the `randomForestSRC` package (Ishwaran et al., 2008). Lastly, in Section S3 we apply the subgroup definitions based on the GBSG analysis to the Rotterdam data [Foekens et al., 2000] which was utilized for "external validation of a Cox prognostic model" [Royston and Altman, 2013].

R code and markdown files for replicating the analyses is available at the GitHub repository: https://github.com/larry-leon/forestSearch. We note that an R package to be published on CRAN is under development. Our code implements parallel computing via the `doFuture` package (Bengtsson, 2021) for the bootstrapping and CV procedures; accordingly the timing of computations depends on the number of available cores.

S1 Power approximations

S1.1 Simulation evaluation of $\hat{\beta} \sim N(\beta, 4/d)$ approximation accuracy

For the random splitting of the FS algorithm we approximate the separate splits via $\hat{\beta}_s^1 \sim N(\beta, 8/d)$ and independently $\hat{\beta}_s^2 \sim N(\beta, 8/d)$. Define $\hat{\beta}_s$ as the Cox model estimate for the subgroup. Since the splits are purely random (the artificially stratified Cox model estimate will be approximately the same as $\hat{\beta}_s$) we have $\hat{\beta}_s^1 + \hat{\beta}_s^2 \approx 2\hat{\beta}_s$.

Figure 1 is based on 20,000 simulations where we generate $N = 60$ subjects from a Cox model with (true) $\beta \approx -0.4$ with number of events (on average) $= 31.9$. Figure 1(a) plots the empirical cdf (ECDF) of the 20,000 estimates corresponding to the first split ($\hat{\beta}_s^1$) along with the $N(\beta, 8/d)$ approximation. Similarly, Figure 1(b) displays the results for the corresponding second split ($\hat{\beta}_s^2$). Figure 1(c) plots the ECDF of 20,000 ($\hat{\beta}_s^1 + \hat{\beta}_s^2$)'s, along with the ECDF of 20,000 $2\hat{\beta}_s$'s. Figure 1(d) plots the ECDF of 20,000 $\min(\hat{\beta}_s^1, \hat{\beta}_s^2)$'s along with the $1-(1-N(\beta, 8/d)^2)$ approximation.

S1.2 Power approximations for censoring rates of 0% and 80%

Figure 2 displays the approximate power under no censoring for scenarios where a subgroup H exists ($n = 60$, 80, or 100) with underlying hazard-ratio for subgroup H varying from 0.5 to 3.0.

For $\theta^{\dagger}(H) \equiv \theta^{\dagger}(ITT) = 0.75$, the approximation is 0.018, 0.009, and 0.004 for $n = 60$, 80, and 100, respectively.

Figure 3 displays the approximate power under heavy censoring of 80% for scenarios where a subgroup H exists ($n = 60$, 80, or 100). Here, for $\theta^{\dagger}(H) \equiv \theta^{\dagger}(ITT) = 0.75$, the approximation is 0.114, 0.096, and 0.081 for $n = 60$, 80, and 100, respectively. However, note that these calculations do not take into account the $d \geq 20$ requirement of the FS procedure which would not be generally met for subgroup sample sizes of 60 and 80, here.

S2 Additional analyses

S2.1 GBSG Analysis

We consider an alternative analysis of the German Breast Cancer Study Group trial data [Schumacher et al., 1994]. In the analysis of the main text we employed the selection criterion where the largest subgroup achieving a consistency rate of at least 90% is chosen based on a broad set of candidate baseline factor cuts. Here we consider the FS criterion of maximizing the consistency rate where lasso is utilized in conjunction with GRF (in a less broad manner in comparison to the main text). Recall the study sample size was $N = 686$ and the observed censoring rate was $\approx 56\%$. The Cox ITT HR estimate (with only treatment as covariate) is 0.69 (0.54, 0.89). There were seven prognostic factors collected: `Estrogen`, `Age`, `Prog`, `Meno`, `Nodes`, `Size`, and `Grade`. The factors `Meno` and `Grade` (`Grade` defined as grade 1/2 vs 3) are categorical and the rest are continuous. The first stage of our algorithm is to apply lasso which selects `Size`, `Grade`, `Nodes`, and `Prog`. Applying GRF (GRF_{60} with a 6-month RMST criterion) selects `Estrogen<=0`. Now, the analyses of Sauerbrei et al. [1999] suggest that `Age` may be prognostic with inflection points at 40 and 55 years (See their Figure 2). We therefore consider these two cuts as candidates ("forced as candidates" regardless of lasso and GRF). In summary, we consider seven candidate factors: `Size<=median(Size)`; `Nodes<=median(Nodes)`; `Prog<=median(Prog)`; `Estrogen<=0`; `Age<=40`; `Age<=55`; and `Grade`. There are then $L = 14$ total single factor subgroups with $L*(L-1)/2+L = 105$ possible subgroups (two-factor combinations): Among these subgroups the number of candidates with sample sizes ≥ 60 and at least 10 events in each arm is reduced to 68.

The FS_{lg} approach, maximizing consistency, estimates $\widehat{H}$ as the subgroup formed by the combination of `Estrogen<=0` and `Prog<=32.5` (The consistency rate is 98.9%). That is, $\widehat{H}$ subjects are those with an estrogen level of 0 and progesterone values at or below 32.5 ($n = 75$). The resulting $\hat{H}$-estimates are $\hat{\theta}(\widehat{H}) = 2.22$ (1.18, 4.2) and (bootstrap bias-corrected) $\hat{\theta}^{*}(\widehat{H}) = 1.59$ (0.82, 3.11). For the complement, $\hat{\theta}(\widehat{H}^{c}) = 0.61$ (0.46, 0.79) and $\hat{\theta}^{*}(\widehat{H}^{c}) = 0.62$ (0.42, 0.9). Figure 4 displays the Kaplan-Meier curves for the estimated subgroups.

To evaluate the quality and stability of the subgroup selection algorithm we apply N-fold and 10-fold cross-validation as described in the main text. Recall for N-fold cross validation we exclude each subject from the analysis and predict their $\widehat{H}$ ($\widehat{H}^{c}$) classification (based on the remaining $n-1$ subjects) where if a subgroup is not identified then the subject is classified as $\widehat{H}^{c}$ (i.e., $\widehat{H} = \emptyset$). In contrast, for 10-fold cross-validation we randomly partition the data into 10 folds and for each fold select $\widehat{H}$ based on the remaining 9 folds to predict the classification for the left out fold. Since this depends on the random partition we repeat this 200 times.

For the N-fold cross-validation, across the $n = 686$ training sets (based on deleting a single subject) there were 622 (91%) subgroups identified wherein `Estrogen<=0` is selected 621 times and `Prog<=32` and `Prog<=33` are selected 342 and 260 times (resp.). Approximately 9% of the subjects are classified as ITT:=$\widehat{H}^{c}$ due to lack of subgroup identification and in total $n = 40$ subjects (N-fold predicted) are classified as $\widehat{H}$ (in contrast to $n = 75$ for the full analysis). Despite the seemingly reasonable agreement between the identified subgroup definitions, the Cox model estimate for the N-fold predicted subgroup $\widehat{H}$ is 0.26 (0.08,0.90) which dramatically differs from the full analysis $\hat{\theta}^{*}(\widehat{H}) = 1.59$ (0.82, 3.11). In addition, across 200 random 10-fold cross-validation analyses the median number of training sets where a subgroup was identified was only 6 out of 10

resulting in a low (median) sensitivity of $sens(\widehat{H}) = 52\%$. That is, among the $\widehat{H}$-classified subjects based on the full analysis the median percentage also $\widehat{H}$-classified in the (10-fold) cross-validation testing samples is approximately only 50%.

In comparison to the analysis of the main text, the FS criterion of maximizing the consistency rate where candidate factors are selected in the above manner does not appear to exhibit preferable stability properties for this data application.

S2.2 ACTG-175 Analysis

We now consider additional analyses of the AIDS Clinical Trials Group Protocol 175 study (Hammer et al., 1996) where, as in the main text, the goal is to identify whether a subgroup exists with a pronounced treatment benefit. Recall we consider the following fifteen baseline covariates: `Age`, `Wtkg`, `Karnof` (Karnofsky score), `Cd40`, `Cd80`, `Hemo` (hemophilia), `HA` (homosexual activity), `Drugs` (history of IV drug use), `Race`, `Gender`, `Oprior` (prior anti-viral therapy), `Symptom`, `Preanti` (days prior antiretroviral therapy), `Str2` (0=naive antiretroviral history,1=experienced), and `Z30` (zidovudine 30 days prior to study).

In contrast to the analysis of the main text, here we only consider cuts for continuous factors at the mean or median. Briefly, the lasso and GRF algorithms lead to the following candidate factors: `Cd40`, `Cd80`, and `Age` are cut at the medians; `Wtkg<=68.04`, `Age<=29`, `Preanti<=406`, `Karnof<=mean`, `HA`, and `Symptom` are also included. Here the mean cut for `Karnof` and median for `Age` were "forced" as candidates. Note that the median for `Karnof` is also the maximum and thus a cut at the median is not viable; And for `Age`, the median (= 34 years) is forced as a candidate since the results of Cui et al. [2023] suggest an inflection point for age between 30 and 40 years (See their Figure 4).

There are then $L = 18$ total single factor subgroups with $L*(L-1)/2+L = 171$ possible subgroups: Among these subgroups the number of candidates with sample sizes ≥ 60 and at least 10 events in each arm is reduced to 151. The largest subgroup with a consistency rate of at least 90% is the subgroup formed by the combination of `Preanti <= 406` and `Age > 34` (98.3%). The resulting $\widehat{Q}$-estimates are $\hat{\theta}(\widehat{Q}) = 0.4$ (0.22, 0.73) and (bootstrap bias-corrected) $\hat{\theta}^*(\widehat{Q}) = 0.48$ (0.27, 0.85). For the complement, $\hat{\theta}(\widehat{Q}^c) = 1.05$ (0.78, 1.4) and $\hat{\theta}^*(\widehat{Q}^c) = 0.98$ (0.67, 1.43). Figure 5 displays the Kaplan-Meier curves for the estimated subgroups.

For the N-fold cross-validation, across the $n = 1083$ training sets there were 1069 (98.7%) subgroups identified wherein the full analysis subgroup $\widehat{Q}$ definitions, `Preanti<=406` and `Age>34`, are reproduced for all (except 1). A total of 14 (1%) subjects were classified as ITT:=$\widehat{Q}^c$ due to lack of subgroup identification and in total $n = 309$ subjects (N-fold predicted) are classified as $\widehat{Q}$ (versus $n = 310$ for the full analysis). The Cox model estimate for the N-fold predicted subgroup $\widehat{Q}$ is 0.45 [95% = 0.25, 0.81] which is similar to the (bootstrap bias-adjusted) full analysis $\hat{\theta}^*(\widehat{Q}) = 0.48$ (0.27, 0.85); And for the complement, the N-fold predicted subgroup $\widehat{Q}^c$ is 1.01 (0.76, 1.36) which is similar to the full analysis $\hat{\theta}^*(\widehat{Q}^c) = 0.98$ (0.67, 1.43).

Across 200 random 10-fold cross-validation analyses the median number of training sets (10 folds) where a subgroup was identified was only 6 out of 10 (The minimum was 3/10 with lower and upper quartiles of 6/10 and 7/10, resp.) resulting in a low (median) sensitivity of $sensCV(\widehat{Q}) = 50\%$. That is, among the $\widehat{Q}$-classified subjects based on the full analysis the median percentage also $\widehat{Q}$-classified in the (10-fold) cross-validation testing samples is approximately only 50%. The median positive predictive value was $ppvCV(\widehat{Q}) \approx 69\%$. For the complement $\widehat{Q}^c$, the medians for *sensCV* and *ppvCV* were 91% and 82%, respectively. In addition, across the 200 random 10-fold cross-validation analyses, the exact $\hat{Q}$ subgroup definition of `Preanti<=406` and `Age>34` was exactly reproduced 20% of the time (median), while definitions based on (cuts of) `Preanti` and `Age` appeared (a median of) 40% and 50% of the time, respectively (and jointly 40%). Although the subgroup estimates are similar to the analysis of the main text, the CV properties are substantially less favorable. We conjecture that this is due to the FS analysis of the main text incorporating a broader set of baseline factor cuts.

The computational timing on an Apple M1 20 core with 69 GB was approximately: 0.2 minutes for the FS analysis; 20 minutes for the 2000 Bootstraps; 11 minutes for the N-fold cross-validation; and ≈ 52 minutes for the 200 random 10-fold cross-validation analyses. In total, the number of minutes was ≈ 84.

S2.3 ACTG-175 Analysis "Added Noise"

We now consider an additional analysis where twenty $N(0,1)$ random variables are artificially considered as baseline factors, along with the actual fifteen clinical factors. There are then 26 continuous factors which we include with cuts at the mean, median, q_1, and q_3. We apply the same criteria as in the analysis of the main text. In this case the GRF approach does not identify any subgroup and hence no factors are added per GRF. The number of factors included was 93: (a) $6*4$ for the 4 cuts applied to the 6 continuous clinical factors; plus (b) $20*3$ cuts for the 20 artificial $N(0,1)$ factors [here median = mean and so only 3 unique cuts per factor]; plus (c) 9 clinical categorical factors. There are then $L = 186$ total single factor subgroups with $L*(L-1)/2+L = 17,391$ possible subgroups: Among these subgroups the number of candidates with sample sizes ≥ 60 and at least 10 events in each arm is reduced to 14,186. The largest subgroup with a consistency rate of at least 90% is the subgroup `Noise11 <= 0` (consistency = 97.7%) which is the 11th $N(0,1)$ factor (Kaplan-Meier subgroup plots displayed in Figure 6). Of course this does not make sense, however we would not know this here pretending these were actual clinical factors. The resulting $\widehat{Q}$-estimates are $\hat{\theta}(\widehat{Q}) = 0.51$ (0.35, 0.76) and (bootstrap bias-corrected) $\hat{\theta}^*(\widehat{Q}) = 0.57$ (0.36, 0.91). For the complement, $\hat{\theta}(\widehat{Q}^c) = 1.33$ (0.92, 1.91) and $\hat{\theta}^*(\widehat{Q}^c) = 1.16$ (0.74, 1.81). Figure 6 displays the Kaplan-Meier curves for the estimated subgroups.

For the N-fold cross-validation, across the $N = 1083$ training sets there were 1065 (98.3%) subgroups identified: The factors `Noise11` and `Preanti` appeared as the first subgroup identifying factor 454 and 602 times, respectively; While a `Noise` factor appeared as a second subgroup identifying factor 643 times. A total of 18 (2%) subjects were classified as ITT:=$\widehat{Q}^c$ due to lack of subgroup identification and in total $n = 615$ subjects (N-fold predicted) are classified as $\widehat{Q}$ (versus $n = 560$ for the full analysis). The Cox model estimates for the N-fold predicted subgroups $\widehat{Q}$ and $\widehat{Q}^c$ are 1.015 (0.71, 1.44) and 0.67 (0.46, 0.99) which dramatically differ from the (bootstrap bias-adjusted) full analysis $\hat{\theta}^*(\widehat{Q}) = 0.57$ (0.36, 0.91) and $\hat{\theta}^*(\widehat{Q}^c) = 1.16$ (0.74, 1.81), respectively. The roles of $\widehat{Q}$ and $\widehat{Q}^c$ seemingly reversed. Across 200 random 10-fold CV analyses the sensitivity and positive predictive value (median) summaries ranged from $65\% - 75\%$. The above N-fold CV discrepancies suggests an underlying instability.

The computational timing on an Apple M1 20 core with 69 GB was approximately: 1.0 minute for the FS analysis; 54 minutes for the 2000 Bootstraps; 30 minutes for the N-fold cross-validation; and ≈ 221 minutes for the 200 random 10-fold cross-validation analyses. In total, the number of minutes was ≈ 305.

S2.3 ACTG-175 Analysis "Added Noise" with lasso

We next consider the previous analysis with the same twenty $N(0,1)$ factors but now with lasso included in the algorithm. As in the previous analysis GRF does not identify a subgroup (GRF is independent of lasso in our implementation), and lasso selects 10 factors (`karnof`, `cd40`, `cd80`, `symptom`, `preanti`, `noise1`, `noise7`, `noise8`, `noise12`, and `noise18`) which are all continuous except `symptom`. We cut the continuous factors at the mean, median, q_1, and q_3 which results in $K = 39$ (non-redundant) binary cuts. There are then $L = 78$ total single factor subgroups with $L*(L-1)/2+L = 3,081$ possible subgroups: Among these subgroups the number of candidates with sample sizes ≥ 60 and at least 10 events in each arm is reduced to 2,591.

The largest subgroup with a consistency rate of at least 90% is the subgroup `Preanti<=744.5` and `Age>34` (consistency = 92.8%) which is identical to the subgroup identified in the main text (i.e., without the twenty noise factors and without using lasso). The resulting $\widehat{Q}$-estimates are

$\hat{\theta}(\widehat{Q}) = 0.52$ (0.32, 0.84] and (bootstrap bias-corrected) $\hat{\theta}^*(\widehat{Q}) = 0.58$ (0.36, 0.96). For the complement, $\hat{\theta}(\widehat{Q}^c) = 1.05$ (0.77, 1.44) and $\hat{\theta}^*(\widehat{Q}^c) = 0.93$ (0.62, 1.39). Note that these are very similar to the estimates in the main text. Figure 7 displays the Kaplan-Meier curves for the estimated subgroups.

For the N-fold cross-validation, across the $N = 1083$ training sets there were 1074 (99.2%) subgroups identified: The factors `Symptom` and `Preanti` appeared as the first subgroup identifying factor 757 and 314 times, respectively; While `Age>34` and a `Noise` factor appeared as a second subgroup identifying factor 314 and 760 times, respectively. A total of 9 (1%) subjects were classified as ITT:=$\widehat{Q}^c$ due to lack of subgroup identification and in total $n = 228$ subjects (N-fold predicted) are classified as $\widehat{Q}$ (versus $n = 382$ for the full analysis). The Cox model estimate for the N-fold predicted subgroups $\widehat{Q}$ and $\widehat{Q}^c$ are 1.91 (1.01, 3.6) and 0.7 (0.52, 0.93) which, as in the last analysis, dramatically differ from the (bootstrap bias-adjusted) full analysis $\hat{\theta}^*(\widehat{Q}) = 0.58$ (0.36, 0.96) and $\hat{\theta}^*(\widehat{Q}^c) = 0.93$ (0.62, 1.39), respectively. Across 200 random 10-fold CV analyses the sensitivity and positive predictive value (median) summaries ranged from $50\% - 74\%$.

In this analysis the incorporation of lasso resulted in identifying a meaningful subgroup, the same as in main text, with corresponding estimates also essentially the same. However, as in the previous analysis (without lasso), the above N-fold CV discrepancies suggests an underlying instability in the presence of including twenty $N(0,1)$ noise factors. We note that the computational timing is similar to the previous analysis.

S2.5 Systolic heart failure data analysis

Analysis evaluating the use of aspirin in the systolic heart failure data (Hsich et al., 2011) available in the `randomForestSRC` (Ishwaran et al., 2008) package ($N = 2,231$ subjects, $p = 38$ baseline covariates, and $K = 78$). We also induce computational challenges by adding 100 noise factors and discuss mitigation approaches when the resulting number of subgroup candidate factors is large, $K = 379$.

We note that this is an observational analysis as the use of aspirin was not randomized. As noted in the discussion section, the FS approach can be extended to adjust for covariates via (stabilized) propensity score weighting Cole and Hernán, 2004, however this is beyond the scope of the current paper and will be a direction for further research. The analyses presented here are for illustration purposes to evaluate the feasibility of including a large number of baseline factors.

The $p = 38$ baseline covariates are summarized in Table 1. Figure 8 displays the Kaplan-Meier curves for the estimated subgroups.

In this example we focus on the feasibility with respect to computational timings (details can be found in the R markdown files at the github site referenced in the introduction). Of the 38 baseline covariates, 13 were continuous and 25 were categorical factors. The number of candidate subgroup (binary) factors was $K = 78$ ($L = 156$ single factor subgroups) with number of all-possible 2 factor combinations $L(L-1)/2 + L = 12,246$. The computational timing on an Apple studio (M1 20 core with 69 GB) was approximately: 0.5 minutes for the FS analysis; 120 minutes for the 2000 bootstraps; 132 minutes for the N-fold cross-validation; and 147 minutes for the 200 random 10-fold cross-validation analyses. In total, the number of minutes was ≈ 399.

We also consider adding 100 random noise factors which resulted in $K = 379$ factors ($L = 758$) and 287,661 all-possible 2 factor combinations. Now, to reduce the computational burden we utilize the variable importance (VI) measure from the `causal_survival_forest` function in the R `grf` package (Tibshirani et al., 2022, Sverdrup et al., 2023). It is not clear what VI values constitute an "important factor"; here we consider excluding factors with VI measures which are less than 10% of the factor with the largest VI measure. That is, if vimax denotes the largest VI measure, then only factors with VI measure $\geq$ vimax/10 are included. This resulted in $K = 151$ and 45,753 all-possible 2 factor combinations. In our simulations we found 400 bootstraps to have good performance for bias-reduction and variance estimation. We therefore applied $B = 500$ bootstraps where the above algorithm (with VI measure criterion) is incorporated in

the procedure. Here the timing for the FS analysis was 1.4 minutes and 56 minutes for the bootstrapping.

Table 1: Summary of events and baseline factors by aspirin use (1=yes, 0=no)

Characteristic	N	0, N = 1,193	1, N = 1,038	Difference	95% CI	p-value
age, Median (IQR)	2,231	54 (45, 61)	56 (50, 63)	-3.1	-4.0, -2.2	0.000
betablok, n (%)	2,231	691 (58%)	738 (71%)	-13%	-17%, -9.2%	0.000
dilver, n (%)	2,231	11 (0.9%)	5 (0.5%)	0.44%	-0.34%, 1.2%	0.328
nifed, n (%)	2,231	46 (3.9%)	53 (5.1%)	-1.3%	-3.1%, 0.57%	0.184
acei, n (%)	2,231	911 (76%)	800 (77%)	-0.71%	-4.3%, 2.9%	0.730
angioten.II, n (%)	2,231	146 (12%)	144 (14%)	-1.6%	-4.5%, 1.3%	0.279
anti.arrhy, n (%)	2,231	284 (24%)	225 (22%)	2.1%	-1.4%, 5.7%	0.252
anti.coag, n (%)	2,231	630 (53%)	269 (26%)	27%	23%, 31%	0.000
aspirin, n (%)	2,231	0 (0%)	1,038 (100%)	-100%	-100%, -100%	0.000
digoxin, n (%)	2,231	912 (76%)	658 (63%)	13%	9.2%, 17%	0.000
nitrates, n (%)	2,231	348 (29%)	391 (38%)	-8.5%	-13%, -4.5%	0.000
vasodilator, n (%)	2,231	81 (6.8%)	55 (5.3%)	1.5%	-0.57%, 3.6%	0.168
diuretic.loop, n (%)	2,231	1,022 (86%)	858 (83%)	3.0%	-0.13%, 6.1%	0.059
diuretic.thiazide, n (%)	2,231	152 (13%)	127 (12%)	0.51%	-2.3%, 3.3%	0.767
diuretic.potassium.spar, n (%)	2,231	327 (27%)	322 (31%)	-3.6%	-7.5%, 0.26%	0.068
lipidrx.statin, n (%)	2,231	305 (26%)	545 (53%)	-27%	-31%, -23%	0.000
insulin, n (%)	2,231	105 (8.8%)	110 (11%)	-1.8%	-4.4%, 0.76%	0.173
surgery.pacemaker, n (%)	2,231	252 (21%)	250 (24%)	-3.0%	-6.5%, 0.61%	0.105
surgery.cabg, n (%)	2,231	201 (17%)	393 (38%)	-21%	-25%, -17%	0.000
surgery.pci, n (%)	2,231	132 (11%)	344 (33%)	-22%	-26%, -19%	0.000
surgery.aicd.implant, n (%)	2,231	284 (24%)	363 (35%)	-11%	-15%, -7.3%	0.000
resting.systolic.bp, Median (IQR)	2,231	108 (96, 120)	110 (100, 122)	-2.8	-4.2, -1.3	0.000
resting.hr, Median (IQR)	2,231	76 (68, 87)	72 (64, 83)	3.6	2.4, 4.7	0.000
smknow, n (%)	2,231	218 (18%)	241 (23%)	-4.9%	-8.4%, -1.5%	0.005
q.wave.mi, n (%)	2,231	88 (7.4%)	191 (18%)	-11%	-14%, -8.1%	0.000
bmi, Median (IQR)	2,231	27.5 (24.3, 31.9)	28.1 (24.8, 32.0)	-0.36	-0.83, 0.11	0.138
niddm, n (%)	2,231	162 (14%)	188 (18%)	-4.5%	-7.7%, -1.4%	0.004
lvef.metabl, Median (IQR)	2,231	20 (15, 25)	20 (15, 25)	-1.0	-1.6, -0.41	0.001
peak.rer, Median (IQR)	2,231	1.10 (1.03, 1.14)	1.10 (1.01, 1.14)	0.01	0.00, 0.02	0.063
peak.vo2, Median (IQR)	2,231	15.9 (12.8, 19.4)	15.6 (12.7, 19.2)	0.23	-0.19, 0.65	0.276
interval, Median (IQR)	2,231	480 (348, 650)	480 (341, 630)	7.9	-10, 26	0.396
cad, n (%)	2,231	319 (27%)	587 (57%)	-30%	-34%, -26%	0.000
died, n (%)	2,231	430 (36%)	296 (29%)	7.5%	3.6%, 11%	0.000
ttodead, Median (IQR)	2,231	5.53 (2.52, 8.26)	4.71 (2.42, 7.34)	0.53	0.28, 0.78	0.000
bun, Median (IQR)	2,231	22 (17, 30)	23 (17, 29)	-0.02	-1.1, 1.0	0.977
sodium, Median (IQR)	2,231	139.8 (138.0, 141.0)	139.8 (138.0, 141.0)	0.09	-0.17, 0.35	0.478
hgb, Median (IQR)	2,231	13.50 (12.80, 14.40)	13.65 (12.85, 14.50)	-0.08	-0.20, 0.03	0.161
glucose, Median (IQR)	2,231	96 (87, 110)	98 (88, 121)	-4.3	-7.9, -0.76	0.017
male, n (%)	2,231	821 (69%)	808 (78%)	-9.0%	-13%, -5.3%	0.000
black, n (%)	2,231	179 (15%)	87 (8.4%)	6.6%	3.9%, 9.3%	0.000
crcl, Median (IQR)	2,231	82 (60, 110)	84 (62, 110)	-1.4	-4.9, 2.2	0.460

[1] Welch Two Sample t-test; Two sample test for equality of proportions

[2] CI = Confidence Interval

S3 GBSG subgroup definitions applied to Rotterdam tumor bank dataset

We apply the subgroup definitions based on the GBSG analysis to the Rotterdam data [Foekens et al., 2000] which was utilized for "external validation of a Cox prognostic model" [Royston and Altman, 2013]. Because the Rotterdam data is purely observational we use (stabilized) propensity-score weighting as described by Cole and Hernán, 2004. Recall, from the GBSG data analysis, subjects without estrogen receptors are anticipated to not benefit from treatment whereas subjects with positive estrogen are anticipated to benefit: The bias corrected hazard ratio estimates were $1.58\ (0.86, 2.9)$ and $0.64\ (0.44, 0.93)$. Though a trend suggesting detriment for subjects without estrogen levels, the 95% CI did not support statistical significance. Figure

9 displays the (stabilized propensity-score weighted) Kaplan-Meier curves along with the corresponding weighted Cox model estimates. The Cox model estimates were 0.55 (0.30, 1.01) and 0.65 (0.49, 0.86), respectively. Here we see that, in contrast to that forecasted via the GBSG results, subjects without estrogen levels trended towards a favorable benefit (Note that there are only 36 [stabilized weighted sum] subjects in the treated arm). Whereas estimates for subjects with positive estrogen levels are fairly consistent with the GBSG forecast.

References

Henrik Bengtsson. A Unifying Framework for Parallel and Distributed Processing in R using Futures. The R Journal, 13(2):208–227, 2021.

Stephen R. Cole and Miguel A. Hernán. Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine, 75(1):45–49, 2004.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. Journal of the Royal Statistical Society Series B: Statistical Methodology, 02 2023. ISSN 1369-7412.

John A. Foekens, Harry A. Peters, Maxime P. Look, Henk Portengen, Manfred Schmitt, Michael D. Kramer, Nils Brünner, Fritz Jänicke, Marion E. Meijer-van Gelder, Sonja C. Henzen-Logmans, Wim L. J. van Putten, and Jan G. M. Klijn. The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients1. Cancer Research, 60 (3):636–643, 02 2000.

Scott M. Hammer, David A. Katzenstein, Michael D. Hughes, Holly Gundacker, Robert T. Schooley, Richard H. Haubrich, W. Keith Henry, Michael M. Lederman, John P. Phair, Manette Niu, Martin S. Hirsch, and Thomas C. Merigan. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine, 335(15):1081–1090, 1996.

Eileen Hsich, Eiran Z. Gorodeski, Eugene H. Blackstone, Hemant Ishwaran, and Michael S. Lauer. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes, 4(1):39–45, 2011.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. Ann. Appl. Statist., 2(3):841–860, 2008. URL https://arXiv.org/abs/0811.1645v1.

Patrick Royston and Douglas Altman. External validation of a cox prognostic model: principles and methods. BMC Medical Research Methodology, 2013.

W. Sauerbrei, P. Royston, H. Bojar, C. Schmoor, and M. Schumacher. Modelling the effects of standard prognostic factors in node-positive breast cancer. german breast cancer study group (gbsg). British journal of cancer, 79(11-12):1752–1760, 1999.

M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, R L Neumann, and H F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. Journal of Clinical Oncology, 12(10):2086–2093, 1994.

Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. policytree: Policy learning via doubly robust empirical welfare maximization over trees. 2023. URL https://CRAN.R-project.org/package=policytree. R package version 1.2.2.

J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager. grf: Generalized random forests. 2022. R package version 2.2.1.

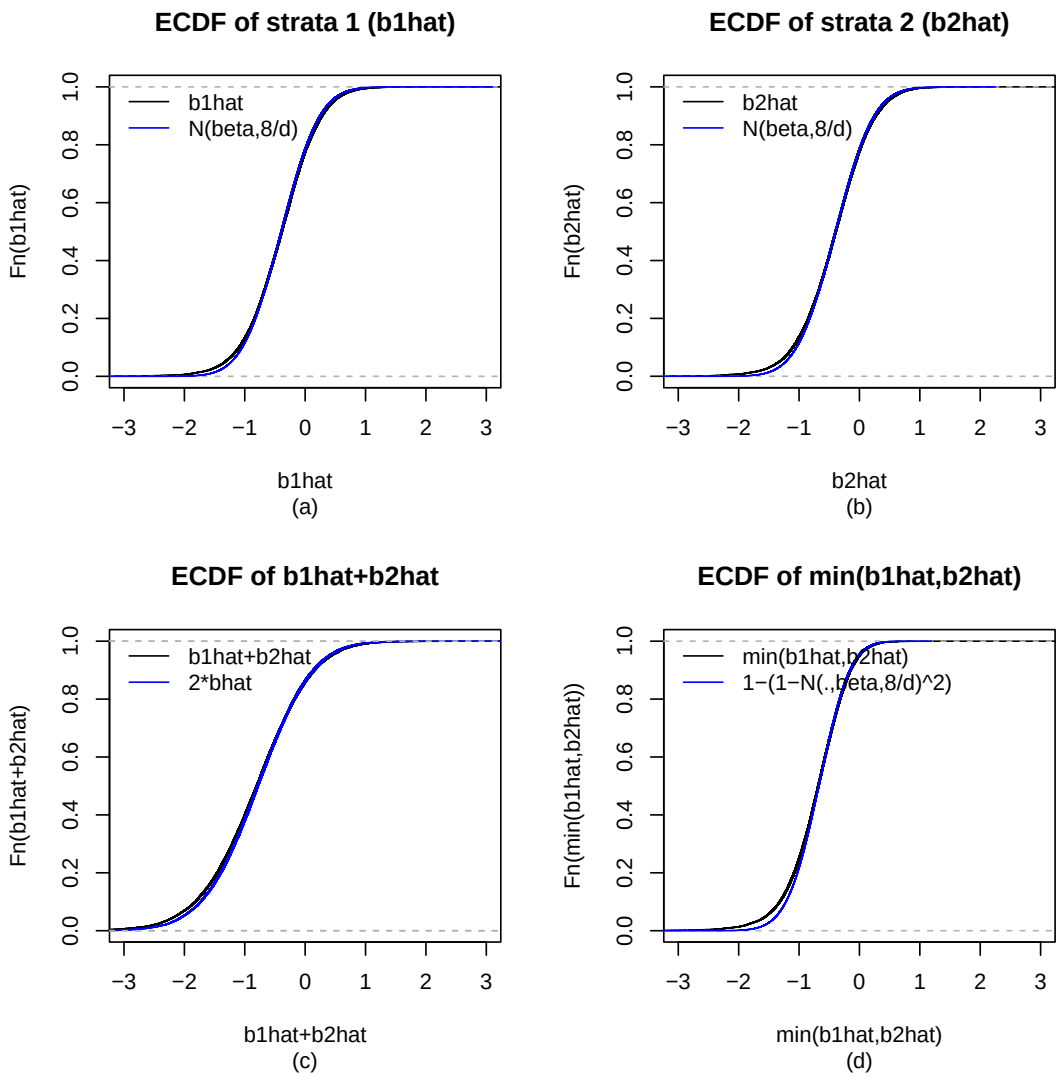

Figure 1: Accuracy of normal approximation $\hat{\beta} \approx N(\beta, 8/d)$

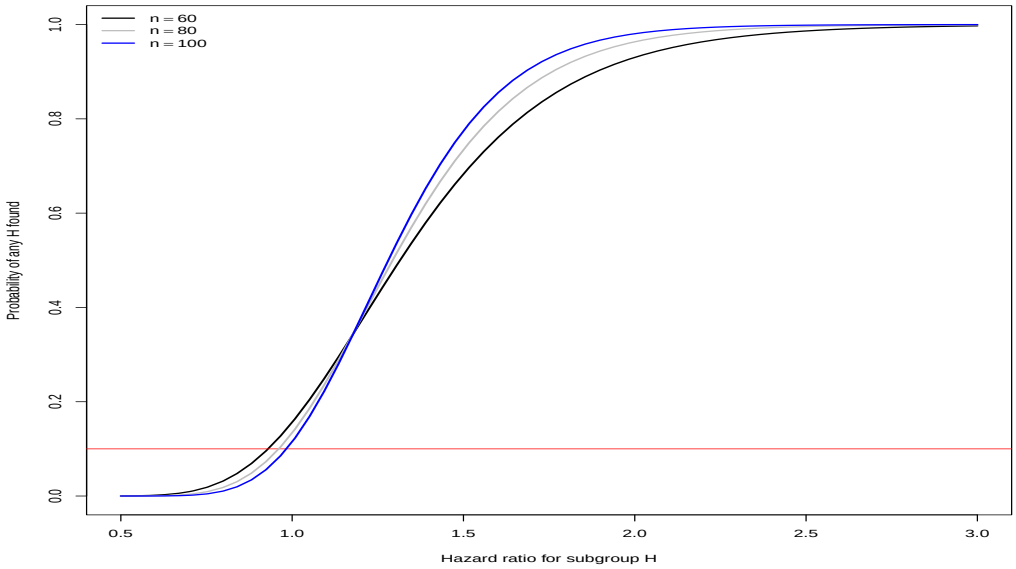

Figure 2: Approximate probability of finding H via FS: Subgroup H of size $n = 60, 80, 100$ exists with underlying hazard-ratio varying from 0.5 to 3.0 and with assumed average censoring rate of 0% so that $d = n$. The horizontal line indicates 10%. Approximately 80% reached at underlying hazard-ratios of 1.69, 1.6, and 1.56, for n = 60, 80, and 100, respectively.

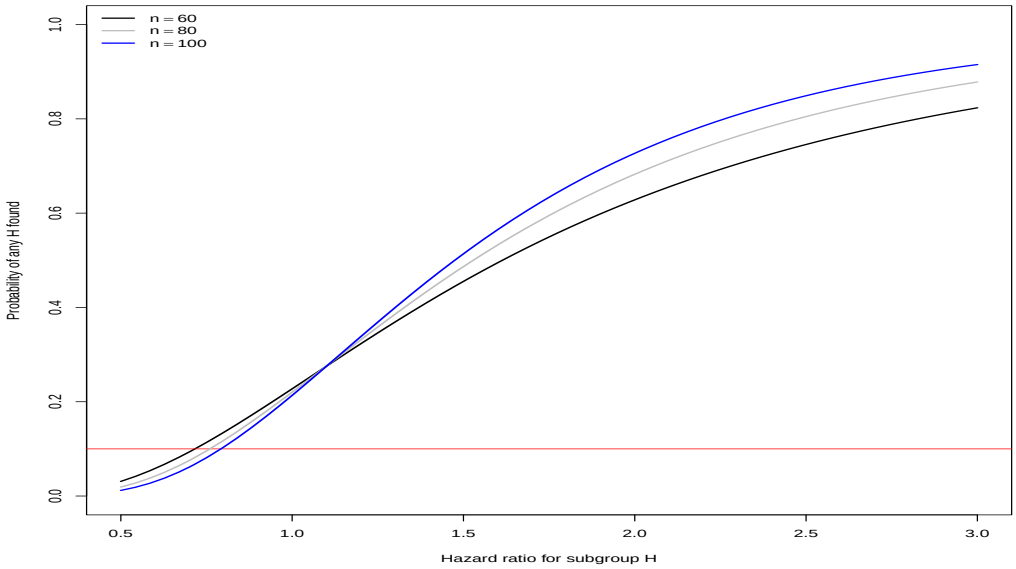

Figure 3: Approximate probability of finding H via FS: Subgroup H of size $n = 60, 80, 100$ exists with underlying hazard-ratio varying from 0.5 to 3.0 and with assumed average censoring rate of 80% so that $d \approx 0.20n$. The horizontal line indicates 10%. Approximately 80% reached at underlying hazard-ratios of 2.83, 2.49, and 2.28, for n = 60, 80, and 100, respectively.

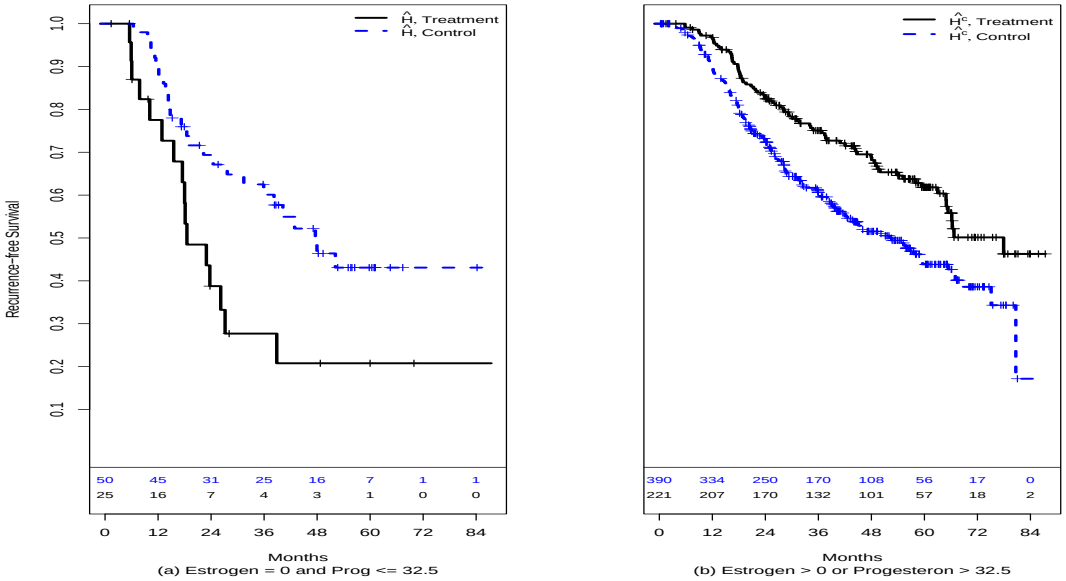

Figure 4: GBSG analysis application of Forest Search. Kaplan-Meier (K-M) curves (un-adjusted for the estimation of subgroups): (a) Forest Search $\widehat{H}$ subgroup treatment estimates; (b) Forest Search $\widehat{H}^c$ subgroup K-M treatment estimates.

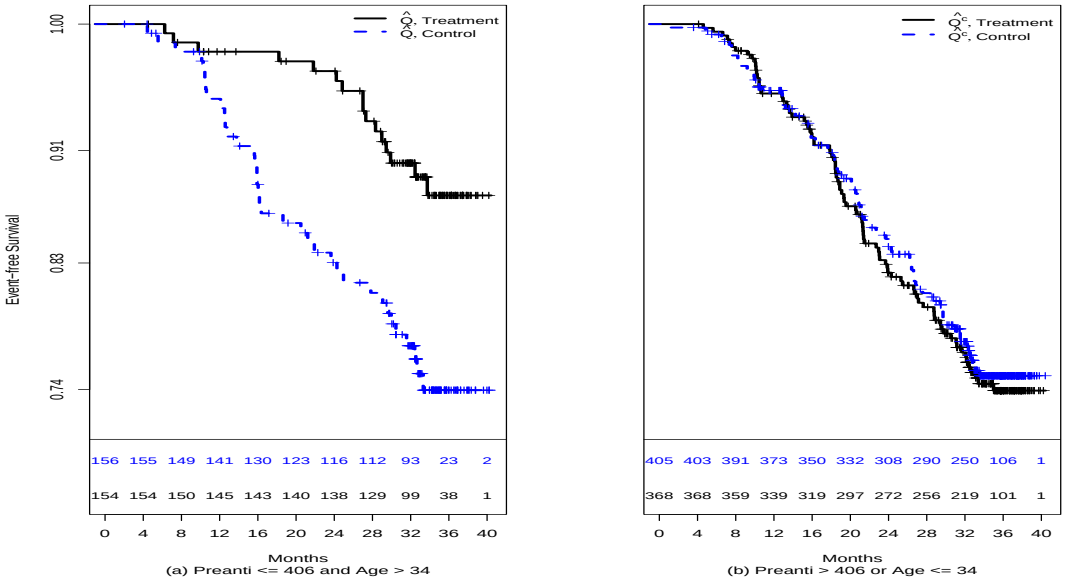

Figure 5: ACTG-175 analysis application of Forest Search. Kaplan-Meier (K-M) curves (un-adjusted for the estimation of subgroups): (a) Forest Search $\widehat{Q}$ subgroup treatment estimates; (b) Forest Search $\widehat{Q}^c$ subgroup K-M treatment estimates.

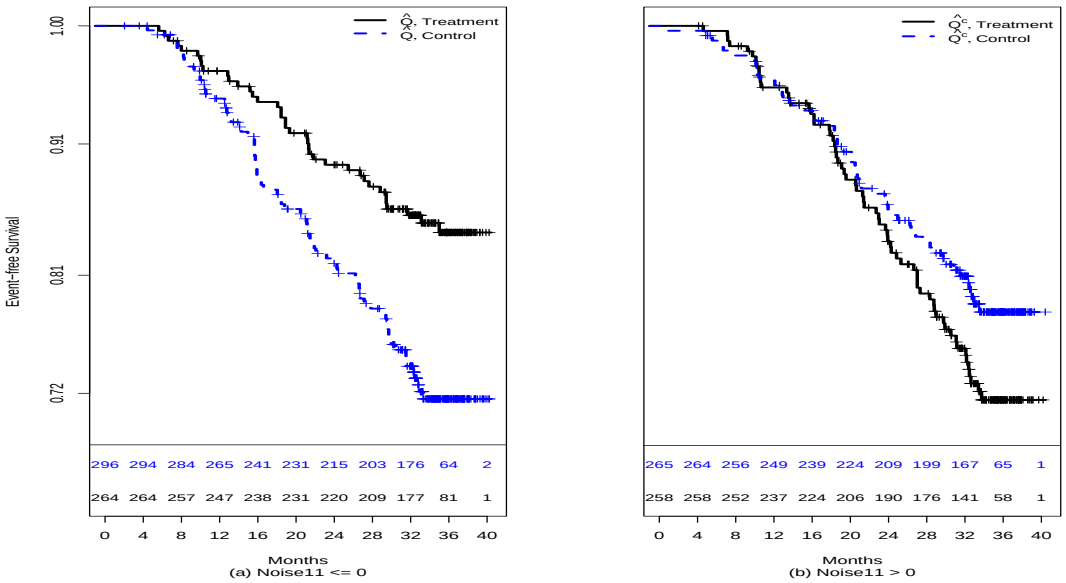

Figure 6: ACTG-175 analysis application of Forest Search. Kaplan-Meier (K-M) curves (unadjusted for the estimation of subgroups): (a) Forest Search $\widehat{Q}$ subgroup treatment estimates; (b) Forest Search $\widehat{Q}^c$ subgroup treatment estimates.

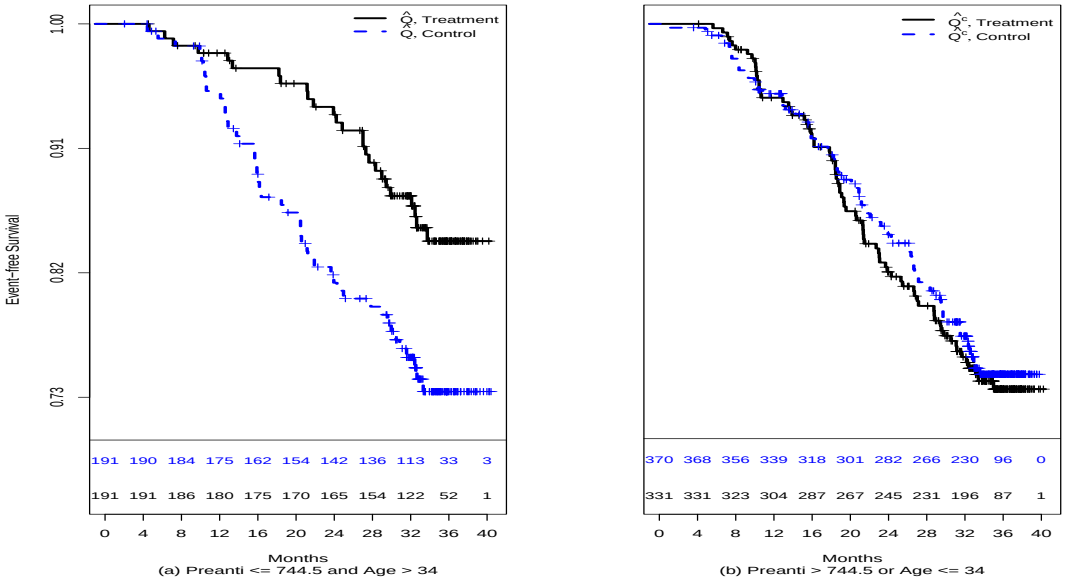

Figure 7: ACTG-175 analysis application of Forest Search. Kaplan-Meier (K-M) curves (unadjusted for the estimation of subgroups): (a) Forest Search $\widehat{Q}$ subgroup treatment estimates; (b) Forest Search $\widehat{Q}^c$ subgroup treatment estimates.

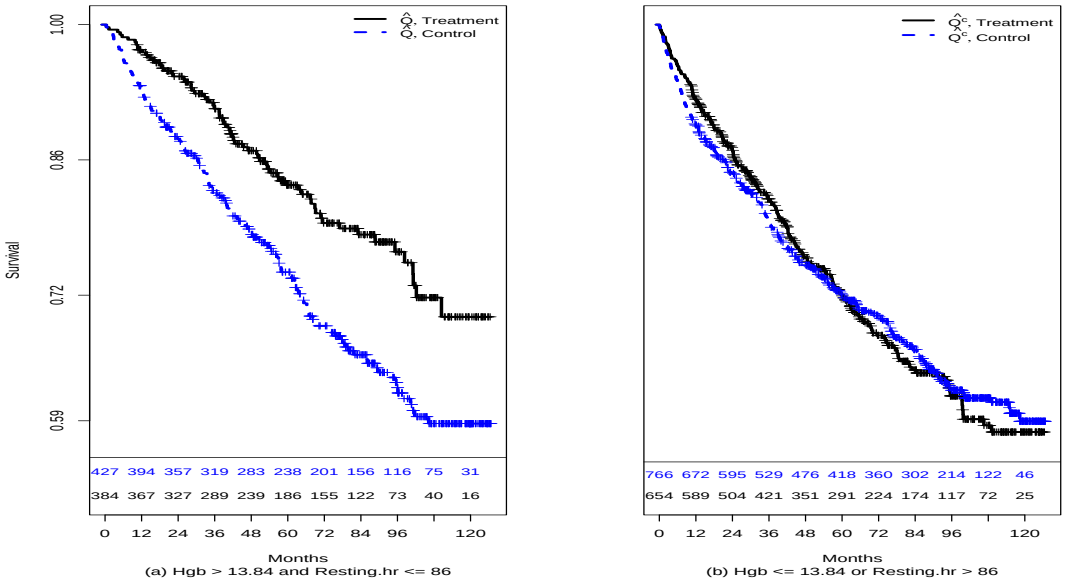

Figure 8: Systolic heart failure data analysis application of Forest Search. Kaplan-Meier (K-M) curves (un-adjusted for the estimation of subgroups): (a) Forest Search $\widehat{Q}$ subgroup treatment estimates; (b) Forest Search $\widehat{Q}^c$ subgroup treatment estimates.

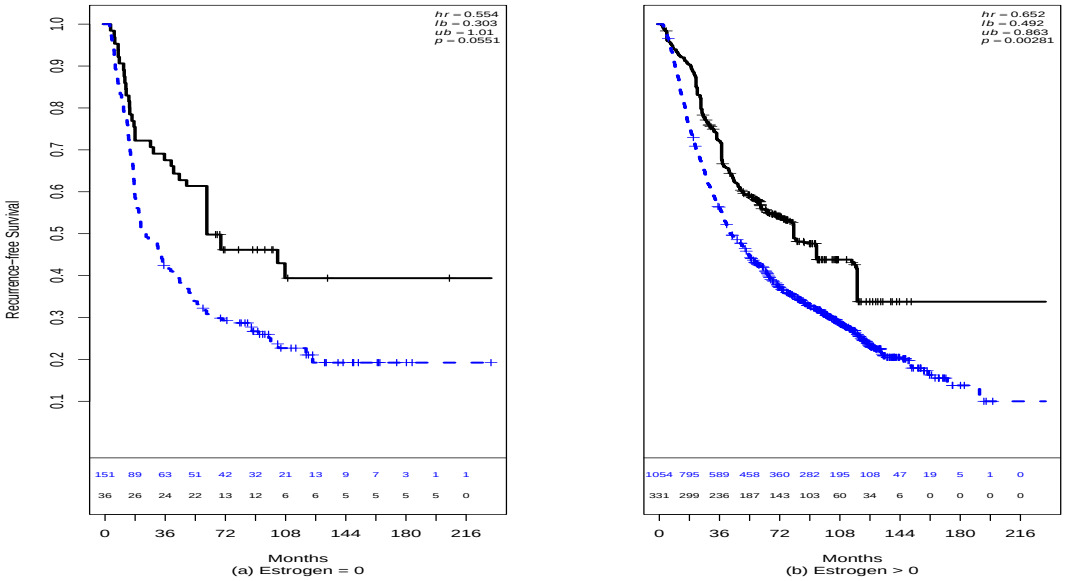

Figure 9: Application of GBSG subgroups to Rotterdam tumor bank data