Received ; | Revised ; | Accepted

DOI: xxx/xxxx

RESEARCH ARTICLE

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

Larry F. León[1] | Thomas Jemielita[1] | Zifang Guo[2] | Rachel Marceau West[1] | Keaven M. Anderson[1]

[1]Biostatistics and Research Decision Sciences, Merck & Co., Inc., New Jersey, USA

[2]Biostatistics, BioNTech SE, New York, USA

Correspondence

*Larry F. León, Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey, USA Email: larry.leon2@Merck.com

Abstract

For survival analysis applications we propose a novel procedure for identifying subgroups with large treatment effects, with focus on subgroups where treatment is potentially detrimental. The methodology is based on extending the idea of all-possible subsets regression from the area of model selection to evaluating all-possible subgroups. The approach, termed forest search, is relatively simple and flexible. Subgroups are screened and selected based on hazard ratio thresholds indicative of harm with assessment according to the standard Cox model. By reversing the role of treatment (switching the treatment indicator) one can seek to identify substantial benefit. We apply a splitting consistency criteria to identify a subgroup considered "maximally consistent with harm". The type-1 error and power can be quickly approximated by numerical integration. To aid inference we describe a bootstrap bias-corrected Cox model estimator with variance estimated by a jacknife approximation. We provide a detailed evaluation of operating characteristics in simulations and compare to virtual twins and generalized random forests where we find the proposal to have favorable performance. In particular, in our simulation setting, we find the proposed approach favorably controls the type-1 error for falsely identifying heterogeneity with higher power and classification accuracy for substantial heterogeneous effects. Two real data applications are provided for publicly available datasets from a clinical trial in oncology and HIV.

KEYWORDS:

Censored data; generalized random forests; virtual twins; bootstrap bias-correction

1 | INTRODUCTION

In oncology trials subgroup analyses via forest plots are standard presentations in regulatory reviews and clinical publications with the goal of evaluating the consistency of treatment effects across the pre-specified subgroups relative to the intention-to-treat (ITT) population. In addition, the European Medicines Agency guideline[1] describes scenarios where there is interest "to identify post-hoc a subgroup where efficacy and risk-benefit is convincing" or "in identifying a subgroup, where a relevant treatment effect and compelling evidence of a favorable risk-benefit profile can be assessed". In a recent review of regulatory considerations for case examples in oncology[2] discuss approvals in the "ITT population despite decreased treatment effect in an important subgroup" as well as approvals in subgroups. The underlying theme in these regulatory reviews was the assessment of an apparent detrimental effect, the evidence for potential harm and biological plausibility.

While pre-specified subgroups provide a higher level of evidence than post-hoc analyses there could be important subgroups based on patient characteristics that are not anticipated or well understood. We investigate approaches for exploratory subgroup identification in survival analysis applications with the goal of identifying an underlying subgroup, H say, consisting of subjects who derive the least benefit from treatment. Ideally subgroup identification would be attempted in Phase 2 in order to inform Phase 3 study design and analysis considerations. In this work we focus on large effects (negative or positive), as "lack of benefit, or mild benefit" may not be sufficient reason to recommend against treatment or to exclude from inclusion in future program development. In the case of an existing detrimental H, the complementary population H^c may potentially be considered to derive benefit with a "higher degree of confidence" relative to the overall ITT population.

The novel methodology in this research, termed forest search (FS), is based on extending the idea of all-possible subsets regression from the area of model selection to evaluating all-possible subgroups formed by combinations of baseline factors. Since the number of combinations can be prohibitively large, we utilize lasso[3] and generalized random forests (GRF)[4,5,6] for pre-selection of candidate factors. For identified subgroups, $\widehat{H}$ and $\widehat{H}^c$ say, inference based on bootstrap bias-corrected estimators is described which accounts for the pre-selection of candidate factors via lasso and GRF. While we are directly targeting identification of H, the primary goal of inference can be with regard to H^c. In addition, by reversing the roles of treatment (switching the treatment indicator) the identification of "harm" can be formulated to identify substantial benefit which will be illustrated in our second real data application.

For identifying H we define (screen) candidates as subgroups with Cox hazard ratio estimates ≥ 1.25 and employ the following *splitting consistency criteria*. Here Cox hazard ratio estimates correspond to the standard model (with only treatment) applied within the subgroup which is common in standard forest plot summaries in oncology trials. Suppose there are subgroups with estimates ≥ 1.25 and for each subgroup we randomly split (e.g, 400 times) the subgroup 50/50 and consider the split consistent with harm if the estimated hazard ratio is ≥ 1.0 for each of the two subgroup splits. We define H-candidates as those with consistency rates at least 90% and define the estimated subgroup, $\widehat{H}$ say, as the subgroup with the highest consistency rate; $\widehat{H}$ is considered "maximally consistent with harm" and $\widehat{H}^c$ is the complement. If no subgroup achieves a consistency rate of at least 90% then define $\widehat{H}$ as null with $\widehat{H}^c$ the ITT population. The consistency criteria heuristically represents –"no matter how you split the subgroup $\widehat{H}$, those splits are both (generally) consistent with harm". The choice of the 1.25 and 1.0 thresholds is based on asymptotic considerations, utilizing the asymptotic normality of the Cox estimator for the subgroup and its random splits to control the type-1 error rate and have reasonable power for large effects.

The *splitting consistency criteria* is similar in spirit to cross-validation, however our goal is not prediction in the general sense of random forest applications, but rather to have independent assessments for evidence of harm which is provided by both (independent) random splits having hazard ratio estimates ≥ 1.0 across repeated sample splitting. Our work is closely related to[7] who consider inference for the largest treatment effect across pre-specified subgroups. However, here the FS algorithm for maximizing the consistency rate does not necessarily correspond to the largest observed treatment effect estimate. Moreover, crucially, we are not pre-specifying a limited set of subgroups but identifying (searching for) subgroups among a large collection of combinations, conceptually a large forest plot. Since we are utilizing GRF (an identification approach itself) in our algorithm for the pre-selection of baseline factors (covariate splits) our procedure can be viewed as a combination of GRF and FS. We refer to[8] for a summary of additional approaches and statistical software.

As a practical illustration we consider a simulated dataset generated as described in Section 3 where we observe 7 baseline factors of which 5 are prognostic ($\{Z_1 - Z_5\}$, say) and the other 2 are non-prognostic ($\{Z_6, Z_7\}$) but correlated with the prognostic factors. In addition, we include 3 additional $N(0, 1)$ random noise variables ($\{Z_8, Z_9, Z_{10}\}$) for a total of 10 baseline factors, $Z_1, Z_2, ..., Z_{10}$. There are 6 binary factors (first 6) and 4 continuous. The underlying subgroup H is an interaction between Z_1 and Z_3 (subjects with $Z_1 = 1$ and $Z_3 = 1$). In this simulated example there are $N = 1000$ subjects, randomized 1:1, where the underlying (marginal) hazard ratio for the harm population H is $\theta^\dagger(H) = 2$ (say), and for the complement H^c, the (marginal) hazard ratio is $\theta^\dagger(H^c) = 0.65$. The number of subjects in the H subgroup is 116 and the number of subjects in the complement equals 884. Figure 1 displays the ITT Kaplan-Meier curves which exhibits a delayed treatment effect pattern (with lack of separation roughly in the first 12 months); The Cox model estimates are 0.73 [95% CI= 0.62, 0.87].

To explore subgroups we proceed as follows. Suppose we cut the continuous variables at the medians so that there are 10 binary factors and $L = 20$ subgroup indicators (further described in section 2). There would then be over 1 million possible subgroup combinations ($2^L - 1$). However, for practical considerations, we only include subgroups formed by a maximum of two factors which is analogous to "tree depths" in random forests (In practice, it may be difficult to clinically interpret subgroups based on 3 or more factors). We also restrict to subgroups with at least 60 subjects (i.e., $\approx n = 30$ per arm under 1:1 randomization) and 10 events in each arm which we consider minimal sample size requirements for Cox and Kaplan-Meier applications. In

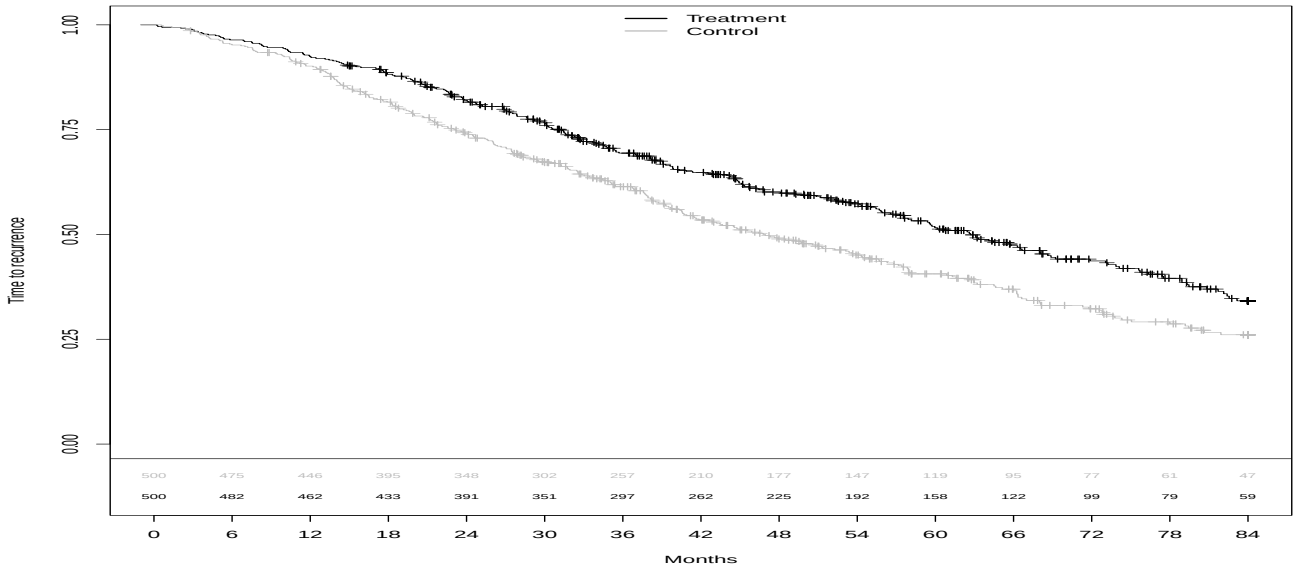

FIGURE 1 Kaplan-Meier curves (ITT) for simulated dataset.

addition, we apply lasso and GRF for selection of continuous variables and splits thereof (described in the following section). Now, applying lasso results in including Z_1, Z_4, Z_5, Z_6, and Z_8, which captures only 3 (Z_1, Z_4, and Z_5 which are binary) of the truly prognostic factors and crucially excludes Z_3 which identifies (along with Z_1) the true subgroup. Applying GRF, Z_1, Z_3 and Z_8 (split at ≤ 0.89) are included as candidates. In this example lasso was somewhat aggressive in excluding factors but the incorporation of GRF re-introduces Z_3. For Z_6 (selected by lasso) this is cut at the median. Both lasso and GRF select the random noise (continuous) factor Z_8 for which we use the cut per GRF ($Z_8 \leq 0.89$). In total FS then evaluates 6 binary factors ($X_1 = Z_1$, $X_2 = Z_3$, $X_3 = Z_4$, $X_4 = Z_5$, $X_5 = (Z_6 \leq \text{med}(Z_6))$, and $X_6 = (Z_8 \leq 0.89)$, say) where the number of all possible combinations is $2^{12} - 1 = 4,095$, of which, only 70 subgroups satisfy the aforementioned criteria.

The resulting estimated H subgroup is the true subgroup and thus Cox model estimates correspond to the oracle estimator where the true subgroup is known a-priori. The Cox estimates, corresponding to FS (and the oracle), are 2.36 [95% CI= 1.53, 3.66] for $\widehat{H}$, and 0.63 [95% CI= 0.52, 0.76] for $\widehat{H}^c$. While these confidence intervals would be valid for the oracle estimator pretending the true subgroup is pre-specified, the FS estimator requires adjustment for the lasso/GRF and searching algorithms. Applying our bootstrap approach the corresponding bias-corrected estimates are 2.04 [95% CI= 1.19, 3.47] and 0.63 [95% CI= 0.48, 0.83], respectively.

This paper is organized as follows. In Section 2 we describe our proposal for subgroup identification along with an asymptotic approximation for the ability to identify (any) H which is the basis for the choice of the FS hazard ratio thresholds (1.25 for screening and 1.0 for consistency). In simulations, Section 3, we compare operating characteristics of the proposed FS approach to virtual twins[9] and generalized random forests[4,5,6] in terms of identification (type-1 error and power) and classification accuracy for correctly identifying subjects in H and H^c. Performance of the bootstrap bias-corrected FS estimators are also evaluated. In Section 4 we analyze two real datasets, the German Breast Cancer Study Group trial data[10] and the ACTG-175 HIV trial[11]. A summary discussion is provided in section 5. Additional details are provided in the supplementary materials.

2 | SUBGROUP IDENTIFICATION APPROACH

We consider the two-sample random censorship model with N observations from a randomized clinical trial (RCT). Let T denote the survival time, C the censoring time, V the treatment assignment, and $Z = (Z_1, Z_2, \ldots, Z_p)$ a p-dimensional collection of baseline covariates. It is of interest to evaluate subgroups formed by combinations of these baseline covariates. We observe the

possibly censored survival time $Y = \min(T, C)$ with $\Delta = I(T \leq C)$ the event indicator. The survival times T are assumed to be independent of C conditional on (V, Z). The observations $(V_i, Z_i, Y_i, \Delta_i)$ for $i = 1, \ldots, N$ are assumed to be iid replicates.

In oncology trials the gold-standard primary ITT analysis is a Cox model with only the treatment arm as a covariate, usually stratified (See for example Amatya[2]). Standard forest plots often proceed by fitting

$$\lambda(t; V) = \lambda_0(t) \exp(\beta V), \tag{1}$$

within the subgroup levels of interest (e.g., by males and females separately).

We assume the candidate subgroups formed by combinations of baseline covariates can be generated by X_k categorical factors based on $Z_1, \ldots, Z_p$ ($k = 1, \ldots, K$ with $K \geq p$). This imposes no restriction on covariates that are naturally categorical, and for continuous covariates lasso and GRF are utilized to pre-select candidate baseline factors. For lasso, selected baseline factors corresponding to non-zero coefficients (per cross-validation) are taken as candidate baseline factors with continuous covariates split at their median. GRF is independently applied to possibly select additional candidate factors with binary cuts corresponding to tree splits. If both procedures select the same continuous factor then the candidate factor per the GRF split is incorporated. Returning to the previous example, the candidate factors were $X_1 = Z_1$, $X_2 = Z_3$, $X_3 = Z_4$, $X_4 = Z_5$, $X_5 = (Z_6 \leq \text{med}(Z_6))$, and $X_6 = (Z_8 \leq 0.89)$; Here lasso selected Z_6 and both selected Z_8, hence $X_5 = (Z_6 \leq \text{med}(Z_6))$ and $X_6 = (Z_8 \leq 0.89)$ were incorporated per our algorithm.

Our procedure for subgroup identification is based on searching through all-possible subgroups formed by candidate factor combinations selected per lasso/GRF. A key restriction that reduces the number of subgroups evaluated is to restrict to subgroups formed by at most 2-factor combinations (similar to "tree depth" in random forests) with a minimum sample size (e.g., of 60 subjects) and with a minimum of number of events (e.g., 10) in each treatment arm. We restrict to 2-factor combinations to simplify clinical interpret-ability and consider (default of) 60 subjects and 10 events in each arm as a reasonable minimal sample size requirement for Cox model applications. The FS procedure for identifying H ($\widehat{H}$) is implemented as follows.

Step 1(a) For candidate baseline factors X_k, $k = 1, \ldots, K$, construct dummy indicators for each unique factor level; Let l_k denote the unique number of values ($k = 1, \ldots, K$) with $L = \sum_{k=1}^{K} l_k$ the number of possible single factor subgroups (e.g., age <= 50, age > 50, gender=Male, gender=Female).

Step 1(b) Let $J_1, \ldots, J_L$ denote the resulting subgroup indicators (For example, $J_1 = 1$ indicates $X_1 = 0$ membership, and $J_2 = 1$ indicates $X_1 = 1$ membership for binary X_1); For all possible combinations of $J_1, \ldots, J_L$ each combination represents a potential subgroup.

Step 2 There are $2^L - 1$ all-possible subgroup combinations. However, we restrict evaluations to subgroup combinations with a minimum subgroup size (default is 60) and with a minimum number of events (default is 10) in each treatment arm. In addition, similar to "tree depths", we limit the number of factors involved in forming combinations to m (default is $m = 2$).

Step 3(a) For subgroup J_{sg} (of size $n_{sg} \geq 60$ and at least 20 events), estimate the Cox model log-hazard ratio $\hat{\beta}_{sg}$ (say), and consider the subgroup as a candidate if $\hat{\beta}_{sg} \geq \log(1.25)$:

Step 3(b) To judge the "consistency with harm", randomly split the J_{sg} subgroup 50/50 and estimate the log-hazard ratio in each of these 2 random splits. Consider this subgroup to be "consistent with harm" if, for each random split, BOTH splits have estimated log-hazard ratios $\geq \log(1.0)$. That is, $\min(\hat{\beta}^1_{sg}, \hat{\beta}^2_{sg}) \geq \log(1.0)$ for log-hazard ratio estimate pairs $\{\hat{\beta}^1_{sg}, \hat{\beta}^2_{sg}\}$ corresponding to each random split;

Step 3(c) Repeat many times (e.g., $R = 400$) to estimate the consistency rate. Let $\{\hat{\beta}^{1r}_{sg}, \hat{\beta}^{2r}_{sg}\}$ denote pairs for the r'th random split for $r = 1, \ldots, R$. The consistency rate is then

$$\hat{p}_{consistency} = \frac{1}{R} \sum_{r=1}^{R} \left\{ I(\min(\hat{\beta}^{1r}_{sg}, \hat{\beta}^{2r}_{sg}) \geq 1.0) \right\}.$$

Step 4 For subgroups with consistency rates at least 90%, choose the subgroup with the highest consistency rate as the estimated H, $\widehat{H}$ ("maximally consistent"); If no subgroup achieves consistency $\geq 90\%$ then consider H as null ($\widehat{H} = \emptyset$).

For the complementary group, H^c is estimated as the complement of $\widehat{H}$, denoted $\widehat{H}^c$; If $\widehat{H}$ is null, then $\widehat{H}^c$ is the ITT population.

In Step 4 the subgroup with the highest consistency rate is chosen, heuristically representing "no matter how you split the subgroup $\widehat{H}$, those splits are (generally) consistent with harm". This puts emphasis on maximizing the consistency rate. To enable additional flexibility, Step 4 can be augmented or modified straightforwardly in several ways: (A) The inclusion of a median threshold, for the experimental arm, the control arm, or both. For example one can restrict to subgroups wherein the experimental arm median is estimable and is below a clinically relevant value (e.g., 3-months); and/or (B) Instead of maximizing the consistency rate, emphasis on larger (or smaller) subgroups can be incorporated by selecting the largest (or smallest) subgroup among those with a high degree of consistency (e.g., at least 90%).

In our first application (Section 4), we apply the maximal consistency criterion as described in Step 4. Our second application seeks to identify the largest subgroup for which there is substantial benefit with an estimated consistency rate ("consistent with benefit") at least 90% (i.e., Step 4 is modified according to (B)). Here we formulate the identification of "harm" in order to identify substantial benefit via switching the treatment roles.

2.1 | Asymptotic Considerations

We now describe how the power for identifying H can be approximated. That is, if a subgroup H exists with underlying (marginal) hazard ratio, denoted $\theta^{\dagger}(H)$, corresponding to harm (e.g., $\theta^{\dagger}(H) = 2.0$) then what is the chance of meeting the consistency criteria? We denote the log-hazard ratio generically by β. Following ideas of Jennison [12] let $L_d(\beta)$ denote the Cox score statistic based on subgroup J_{sg} of Step 3 with a total number of d observed events (with sample size $n_{sg} \geq 60$ and $d \geq 20$) and corresponding log-hazard ratio estimate $\hat{\beta}_{sg} \geq \log(1.25)$ (according to Step 3).

For the random splitting step (Step 4) of the FS algorithm form $\tilde{L}_d(\beta) = L_{d_1}(\beta) + L_{d_2}(\beta)$ where $L_{d_1}(\beta)$ and $L_{d_2}(\beta)$ are based on randomly generating an (artificial) stratification factor ($\sim$ Bin(0.5)) with $\tilde{L}_d(\beta)$ the Cox score statistic based on the artificial stratification. Denote the Cox model estimates based on the above random splits by $\hat{\beta}^1_{sg}$, and $\hat{\beta}^2_{sg}$, respectively. Due to the purely random splitting $\hat{\beta}_{sg} \approx \tilde{\beta}_{sg}$ where $\tilde{\beta}_{sg}$ is the (randomly) stratified Cox estimate. Applying the normal approximation [12] for the log-hazard ratio we have $\hat{\beta}_{sg}$ is approximated by $(4/d)\tilde{L}_d(0)$, which in turn is approximated in distribution by a $N(\beta, 4/d)$ random variable [12]. Similarly, $\hat{\beta}^1_{sg}$ and $\hat{\beta}^2_{sg}$ are each approximated (independently) by $(8/d)L_{d_1}(0) \approx N(\beta, 8/d)$, and $(8/d)L_{d_2}(0) \approx N(\beta, 8/d)$, since for both random splits $d_1 \approx d_2 \approx d/2$. Write these approximations as $L_{d_1}(0) \approx (d/8)\hat{\beta}^1_{sg}$, $L_{d_2}(0) \approx (d/8)\hat{\beta}^2_{sg}$, and $\tilde{L}_d(0) \approx (d/4)\hat{\beta}_{sg}$. Now, by construction $\tilde{L}_d(0) = L_{d_1}(0) + L_{d_2}(0)$ and we thus have, approximately

$$\hat{\beta}_{sg} \geq \log(1.25) \iff \hat{\beta}^1_{sg} + \hat{\beta}^2_{sg} \geq 2 * \log(1.25). \tag{2}$$

For a subgroup H with underlying log-hazard ratio β we can approximate the probability of identifying H via $P(W_1 + W_2 \geq 2 * \log(1.25), \min(W_1, W_2) \geq \log(1.0)) =$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(w_1 + w_2 \geq c_1) I(w_1 \geq c_2) I(w_2 \geq c_2) \varphi(w_1; \beta, 8/d) \varphi(w_2; \beta, 8/d) dw_1 dw_2, \tag{3}$$

where $c_1 = 2\log(1.25)$, $c_2 = \log(1.0)$, $\{W_1, W_2\} \sim N(\beta, 8/d)$ (ind.), and $\varphi(\cdot; \beta, 8/d)$ denotes the normal density with mean β and variance $8/d$.

Figure 2 displays (3) for scenarios where a subgroup H exists ($n_H = 60, 80$, or 100) with underlying hazard-ratio for subgroup H varying form 0.5 to 3.0. Of course, hazard ratios ≤ 1.0 correspond to non-negative treatment effects and the horizontal line is at 10% suggesting the "type-1" error rate is reasonable. The "power" also seems reasonable, generally $\geq 70\%$ for identifying underlying hazard ratios in the ≥ 2.0 range. Our choice of the 1.25 and 1.0 thresholds was based on the desire to control the rate for finding a subgroup H to be below 10% when the underlying hazard ratio for H is below 1.0, where treatment is strictly non-negative in terms of benefit. Moreover, if the treatment effect is uniform and beneficial then for a random subgroup H, Cox model estimates will randomly fluctuate around the ITT effect ($\theta^{\dagger}(H) \approx \theta^{\dagger}(ITT)$, say). For example, for $\theta^{\dagger}(H) := \theta^{\dagger}(ITT) = 0.75$, the above approximation is 0.049, 0.033, and 0.022 (for $n_H = 60$, 80, and 100, respectively) indicating reasonable control of type-1 error. In simulations we evaluate the type-1 error (for falsely identifying a non-existent H) and power under various scenarios designed to mimic potential Phase 2 and Phase 3 trial conditions.

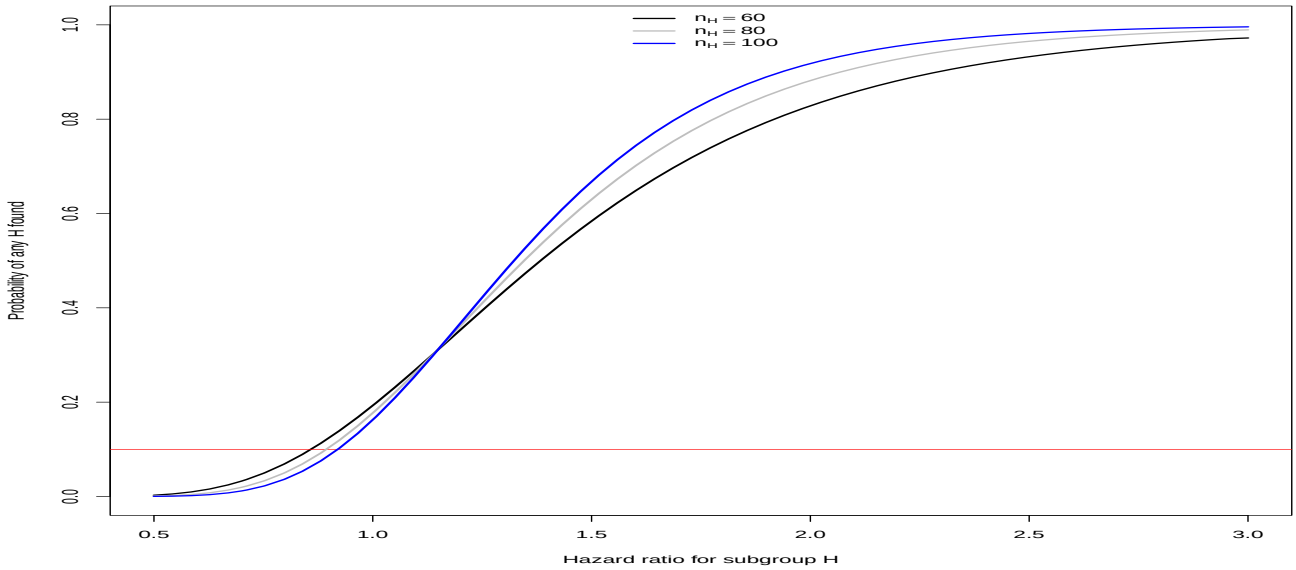

FIGURE 2 Approximate probability of finding H via FS: Subgroup H of size $n_H = 60, 80, 100$ exists with underlying hazard-ratio $\theta^{\dagger}(H)$ varying from 0.5 to 3.0 and with assumed average censoring rate of 45% so that $d \approx 0.55 \times n_H$. The horizontal line indicates 10%. Approximately 80% reached at underlying hazard-ratios of 1.9, 1.8, and 1.7, for $n_H = 60$, 80, and 100, respectively.

3 | SIMULATIONS

Our simulation setting is based on the German Breast Cancer Study Group trial data (GBSG)[10] that is available in the R statistical software[13] survival library gbsg[14]. The study sample size was $N = 686$ and the outcome of interest was tumor recurrence following the addition of hormonal therapy (yes/no) in the adjuvant setting. The observed censoring rate was $\approx$ 56%. The dataset consists of a mixture of randomized and non-randomized subjects from the trial and seven baseline factors are available including estrogen receptors (fmol/l), age (years), progesterone receptors (fmol/l), menopausal status (post vs pre), number of positive lymph nodes, tumor size (mm), and tumor grade (grade 1/2 vs 3). We denote these by W_1 (`Estrogen`), W_2 (`Age`), W_3 (`Meno`), W_4 (`Progesterone`), W_5 (`Nodes`), W_6=(`Size`), and W_7 (`Grade`), respectively. A detailed evaluation of these prognostic factors is described by Sauerbrei[15].

In order to mimic a randomized clinical trial and to have the flexibility to simulate our desired sample sizes, we first randomly draw treatment arms from the gbsg dataset, retaining the observed covariates, for a large "super-population" of 5,000 subjects. Specifically, for two synthetic treatment arms, 2,500 subjects are randomly drawn with replacement (for each arm) from the $N = 686$ subjects to construct a large population that mimics the covariate structure of the dataset. Simulations are then based on randomly sampling from this "super-population".

In the sequel we define $\theta^{\dagger}(G)$ as the marginal hazard ratio for a generic subgroup G of the "super-population" which corresponds to the (un-adjusted) Cox model, based on subgroup G within the "super-population" with only the treatment arm as a covariate. For ease of notation denote the indicator variable defining binary cuts generically as $I(U \leq \cdot) = (U \leq \cdot)$.

To generate outcomes we consider a true model depending on prognostic baseline factors $Z_1 - Z_5$, which we take as observable, where the H subgroup is generated by a treatment interaction between Z_1 (defined below) and Z_3, with Z_3 denoting post-menopausal status ($Z_3 = W_3$). The remaining prognostic factors are $Z_2 = (W_2 \leq \text{med}(W_2))$, $Z_4 = (W_4 \leq \text{med}(W_4))$, and $Z_5 = (W_5 \leq \text{med}(W_5))$. In addition, $Z_6 = W_6$, and $Z_7 = W_7$ are observed but non-prognostic, though correlated with $Z_1 - Z_5$ per the gbsg dataset. Now, for Z_1 we define k_{p_H} such that for $Z_1 = (W_1 \leq k_{p_H})$, the proportion of subjects in the "super-population" subgroup with $Z_1 = 1$ and $Z_3 = 1$ is $\approx p_H$.

The true outcome generating model is a Weibull structure

$$\log(T) = \mu + \beta_0 \mathrm{V} + \beta_1 \mathrm{V} Z_1 Z_3 + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 Z_4 + \beta_6 Z_5 + \tau\epsilon, \tag{4}$$

with V denoting treatment, ϵ is from the standard extreme value distribution and τ is a dispersion parameter. The interaction between Z_1 and Z_3 represents the subgroup H= $\{Z_1 = 1\} \cap \{Z_3 = 1\}$ with an underlying proportion of subjects $\approx p_H$. The analyst has available factors Z_1-Z_5 plus 2 non-prognostic factors Z_6 and Z_7. The parameters β_0 and β_1 determine the treatment effects where $\beta_1 = 0$ corresponds to no subgroup effect ($H = \emptyset$). For a simulated trial of size N, the average number of subjects in the H subgroup is $N \times p_H$ and the average number of subjects in the complement $H^c = \{Z_1 = 0\} \cup \{Z_3 = 0\}$ is $N \times (1 - p_H)$. For example, we consider a scenario with $p_H \approx 13\%$ where the size of H is relatively small, but practically important and presents a challenge to identify H and to the interpretation of an overall (ITT) treatment effect.

Writing the above data-generating model as

$$\log(T) = \mu + \beta_0 \mathrm{V} + \beta_1 \mathrm{V} Z_1 Z_3 + \boldsymbol{\beta_2}' \boldsymbol{Z}_2 + \tau\epsilon, \text{ say}, \tag{5}$$

we denote the corresponding hazard function when treatment is set to v (0 under control; 1 under treatment) for subjects with given prognostic values ($\boldsymbol{Z} = \boldsymbol{z}$, say) as

$$\lambda_v(t; \boldsymbol{z}) = \lambda_0(t) \exp(\gamma_0 \mathrm{v} + \gamma_1 \mathrm{v} z_1 z_3 + \boldsymbol{\gamma_2}' \boldsymbol{z}_2), \tag{6}$$

where $\gamma = -\tau\beta$ [16], say. The parameters μ, $\boldsymbol{\beta}_2$, and τ are based on Weibull model fits to the observed gbsg data with β_0 and β_1 then chosen to generate (marginal) hazard ratio subgroup effects of interest in the "super-population" (e.g., $\theta^\dagger(H) = 2.0$, and $\theta^\dagger(H^c) = 0.65$). Censoring is also generated by a Weibull model based on the observed data in order to have approximately 46% censoring.

We evaluate the operating characteristics for identifying and estimating H and H^c under various sample sizes and treatment effects. Under the null model with $\beta_1 = 0$ ($H = \emptyset$), H^c is the ITT population with (marginal) hazard ratio $\theta^\dagger(ITT)$. We note for fixed p_H as $\beta_1 \neq 0$ varies, inducing subgroup effect $\theta^\dagger(H)$, the overall (ITT) population effect $\theta^\dagger(ITT)$ will also vary except that the subgroup effect for H^c will remain constant.

For identification and estimation we are targeting marginal hazard ratios for H and H^c in the "super-population" ($\theta^\dagger(H)$ and $\theta^\dagger(H^c)$, resp.), where subgroup analyses are based on un-adjusted Cox models. As described by Aalen [16] the marginal effects for H and H^c will generally differ from their *controlled direct effects* which we denote by $\theta^\ddagger(H)$ and $\theta^\ddagger(H^c)$. We note from (6) that $\theta^\ddagger(H) = \exp(\gamma_0 + \gamma_1)$, and $\theta^\ddagger(H^c) = \exp(\gamma_0)$. When describing FS estimation properties in Section 3.2 we will consider accuracy in terms of both $\theta^\dagger(.)$ and $\theta^\ddagger(\cdot)$.

Now, in addition to the proposed FS approach we evaluate virtual twins [9] and GRF [4,5,6] procedures for subgroup identification. For the virtual twins approach, to account for censoring we employ a "censoring unbiased transformation" [17] (See Steingrimsson [18] for a double-robust implementation). Virtual twins is implemented via the R package `aVirtualTwins` [19], and GRF is implemented using the `causal_survival_forest` function in the R `grf` package [20].

For each approach we restrict to subgroups with sample sizes of at least 60 subjects, and restrict to subgroups based on combinations of at most 2 baseline factors (i.e., tree depths of 2). For GRF and VT we consider the following.

GRF: GRF targets RMST and we denote GRF as RMST based on the truncation point $\tau = \min(\tau_0, \tau_1)$ where τ_0 and τ_1 are the largest non-censored (event) outcomes for the control and treatment groups (respectively). An RMST benefit of (at least) 6 months for control is required for selection of a subgroup H, where among tree depths of 1 and 2, the subgroup with the largest RMST benefit ($\geq$ 6 months in favor of control) is selected.

GRF.60: The GRF procedure employs a double-robust approach for estimating RMST that involves estimation of the censoring distribution. As such, the choice of the truncation point can be influential. To reduce the potential instability we consider GRF.60 which uses $\tau_{60} := 0.6 \min(\tau_0, \tau_1)$.

VT(24): We consider the virtual twins approach targeting survival rates at $t = 24$ months. A treatment effect of $\delta \geq 0.225$, in favor of control, is required for selection of H.

VT(36): Same as VT(24) but with $t = 36$.

To quantify the classification properties we consider the following sensitivity and positive predictive value measures. For estimated subgroup $\widehat{H}$ define $sens(\widehat{H})$ and $ppv(\widehat{H})$ as

$$sens(\widehat{H}) = \#\{i \in \widehat{H} \cap H\}/\#\{i \in H\}, \text{ and } ppv(\widehat{H}) = \#\{i \in \widehat{H} \cap H\}/\#\{i \in \widehat{H}\},$$

with measures for the complement $\widehat{H}^c$ defined analogously. Note that there always exists $\widehat{H}^c$ for any procedure, since if a candidate subgroup does not meet the criteria of a procedure then $\widehat{H} = \emptyset$ and the estimated complement is set to the overall ITT population ($\widehat{H}^c = \Omega$, say). Under the null when no subgroup H exists, the denominator in $sens(\widehat{H})$ is zero and the numerator in $ppv(\widehat{H})$ is zero, thus $sens(\widehat{H})$ is undefined and $ppv(\widehat{H}) = 0$.

3.1 | Chance of Finding Any Subgroup H

In our simulation study we consider three data generation model scenarios, models M_1, M_2, and M_3, where performance of the methods are evaluated under "null' and alternative subgroup effect conditions across 20,000 simulations. For each scenario we consider the performance when the factors $Z_1 - Z_7$ are observed (recall $Z_1 - Z_5$ are prognostic, whereas Z_6 and Z_7 are non-prognostic but correlated per the gbsg dataset), as well as when additional noise factors are included which are represented as independent standard normal random variables (e.g., $Z_8 \sim N(0,1)$, $Z_9 \sim N(0,1)$, and $Z_{10} \sim N(0,1)$, each independently).

Table 1 displays the probabilities for identifying a subgroup H (denoted $any(H)$) as well as the classification rates for each analysis approach under the null ($H = \emptyset$) and alternative, where the (marginal) hazard ratio for the subgroup H is $\theta^\dagger(H) = 2.0$ under each model $M_1 - M_3$.

Under the null, we consider rates above 10% for falsely identifying a subgroup H (type-1 error) as generally inflated. In Table 1 cases are bold-faced where under the null $any(H) \geq 0.10$.

Under model M_1, the first block in Table 1 , there are $N = 700$ subjects where under the null ($H = \emptyset$, denoted "M_1 Null") and alternative (denoted "M_1 alt") the hazard ratios for the ITT population, $\theta^\dagger(ITT)$, are similar at 0.7, and 0.71, respectively. Under the alternative the proportion of subjects in the true H subgroup is $p_H \approx 13\%$ and the hazard ratio for H^c is $\theta^\dagger(H^c) = 0.65$. In the scenario when only factors $Z_1, \ldots, Z_7$ are included in the analysis (The first 6 columns), under the null, all approaches control the type-1 error at $\leq 5\%$ except GRF which is at 25%. Under the alternative, FS_l and FS_{lg} both outperform GRF_{60} (and virtual twins) with higher rates for identifying H and classification accuracy. For example, for FS_{lg} the chance of identifying any H is 86% and the accuracy for correctly classifying subjects in H, $sens(\widehat{H})$, is 82%; Whereas for GRF_{60} these rates are 72% and 66%, respectively. When the analysis includes three additional random noise factors, columns 7-12, the type-1 error rates for the GRF approaches are both quite elevated (61%, and 27% for GRF and GRF_{60}, resp.) with FS_{lg} slightly elevated at 11% whereas FS_l is at 2%. Moreover, despite the higher type-1 error for GRF_{60}, FS_l and FS_{lg} both have higher classification accuracy (e.g., $sens(\widehat{H})$ is 64% [74%] for FS_l [FS_{lg}] compared to 52% for GRF_{60}).

A similar pattern to model M_1 is found under M_2 where we consider a smaller sample size but with a higher proportion of subjects in the H subgroup. Specifically, under model M_2 there are $N = 500$ subjects where under the null and alternative $\theta^\dagger(ITT)$ is 0.69 and 0.79 (resp.), while under the alternative $p_H \approx 20\%$ and $\theta^\dagger(H^c) = 0.69$. In addition, we consider the performance when five additional random noise factors are included in the analysis. The type-1 errors are similar to model M_1, however the identification and accuracy rates are higher for H relative to model M_1 even though the incidence rate for H is only moderately increased (The average size for H under models M_1 and M_2 are 89 and 101, resp.).

Lastly, we consider a (relatively) small sample size of $N = 300$ in model M_3 with a strong ITT treatment effect under the null where $\theta^\dagger(ITT) = 0.55$. In this scenario all the approaches control the type-1 error rates below 5% except for GRF and GRF_{60} which are slightly elevated (13%, and 7%, resp.) when five additional random noise factors are included in the analysis. In this scenario GRF has the strongest performance, albeit with the aforementioned increased type-1 error rate, whereas FS_{lg} has the highest accuracy while maintaining the type-1 error at $\leq 2\%$.

We note that while the accuracy for classification of H^c subjects via $sens(\widehat{H}^c)$ remains seemingly high in the presence of additional noise factors ($\geq 87\%$), the FS_{lg} approach is around 7% higher compared to GRF_{60} for some scenarios (e.g., 96% vs. 89% under the M_2 alternative). Though not dramatic, this could be important in clinical practice from an individual patients' perspective.

In this simulation setting, when random noise factors are included in the analyses the GRF approach is more susceptible to falsely identifying subgroups especially under M_1 and M_2. Intuitively, with the addition of noise factors there is more opportunity to "randomly form erroneous splits". For virtual twins, the type-1 errors are not materially increased but the accuracy

performance is generally diluted across the scenarios. The FS_l approach is the most stable with a slight decrease in performance, while FS_{lg} inherits an increased type-1 error by the utilization of GRF_{60}, but to a much lesser extent than GRF_{60} itself. In contrast, under M_3, when there is the strongest ITT treatment effect under the null, the type-1 errors for GRF are dramatically decreased relative to M_1 and M_2 (From $\approx 60\%$ for GRF under M_1 and M_2 to 13% under M_3.). We conjecture that this is due to the GRF selection criteria which requires an estimated 6-month benefit in favor of control, which is less likely with a more pronounced ITT treatment effect (Note that under the nulls of $M_1 - M_3$ the ITT treatment differences with respect to RMST are $\approx$ 7.2, 7.4, and 11.5 months, resp.). Generally, for each approach under the null, the chance of forming subgroups with an estimated benefit randomly in favor of control is less likely the stronger the ITT treatment effect benefit. Table 1 also provides the power approximation (3) for the FS procedure under the null and alternative for models $M_1 - M_3$ which appears reasonably accurate (see footnotes a-f).

3.2 | FS Bootstrap Bias-Correction and Variance Estimation

By the nature of the FS procedure, we expect (raw) un-adjusted Cox model point estimates based on $\widehat{H}$ to be upwardly biased due to the hazard ratio thresholds, especially for $\theta^{\dagger}(H) \leq 1.25$ (Since by construction raw point estimates are ≥ 1.25 for $\widehat{H}$). However the bias can also be pressured in the opposite direction depending on the proportion of H^c subjects (incorrectly) included in $\widehat{H}$ and the value of $\theta^{\dagger}(H)$ relative to $\theta^{\dagger}(H^c)$ (e.g., mixture of $\theta^{\dagger}(H) = 2.0$ vs. $\theta^{\dagger}(H^c) = 0.65$). In general, there is potential for exacerbating estimation bias due to the subgroup selection. For bias-correction, we proceed on the Cox regression coefficient scale, denoted $\hat{\beta}(\widehat{H})$, and then exponentiate to obtain point estimates and confidence intervals for hazard ratios $\hat{\theta}(\widehat{H}) := \exp(\hat{\beta}(\widehat{H}))$.

The bias corrected estimator $\hat{\beta}^*(\widehat{H})$, described below, is along the lines of Harrell[21]. For the observed data, with estimated subgroup $\widehat{H}$, define $\hat{\beta}(\widehat{H})$ as the estimated Cox model regression parameter. Analogously, for bootstrap samples $b = 1, \ldots, B$ (e.g., $B = 500$) with estimated subgroup $\widehat{H}^*_b$, let $\hat{\beta}^*_b(\widehat{H}^*_b)$ denote the estimated Cox model parameter for the bootstrap sample based on subgroup $\widehat{H}^*_b$. In addition, let $\hat{\beta}(\widehat{H}^*_b)$ denote the Cox model parameter for the observed data based on the bootstrap estimated subgroup $\widehat{H}^*_b$ (That is, the Cox model estimate applied to the observed data within the subgroup defined by $\widehat{H}^*_b$). Define $\hat{\beta}^*_b(\widehat{H})$ similarly and form the bias terms for $\hat{\beta}(\widehat{H})$:

$$\eta^*_b(\widehat{H}^*_b) = \hat{\beta}^*_b(\widehat{H}^*_b) - \hat{\beta}(\widehat{H}^*_b), \text{ and } \eta^*_b(\widehat{H}) = \hat{\beta}^*_b(\widehat{H}) - \hat{\beta}(\widehat{H}).$$

Correspondingly, for the complementary subgroup, define $\eta^*_b(\widehat{H}^{c*}_b)$ and $\eta^*_b(\widehat{H}^c)$ for $\hat{\beta}(\widehat{H}^c)$ analogously. Let $\{(\eta^*_b(\widehat{H}^*_b) + \eta^*_b(\widehat{H})), (\eta^*_b(\widehat{H}^{c*}_b) + \eta^*_b(\widehat{H}^c))\}$ denote bootstrap samples $b = 1, \ldots, B$. The bias-corrected estimators are defined as

$$\hat{\beta}^*(\widehat{H}) = \hat{\beta}(\widehat{H}) - (1/B)\sum_{b=1}^{B}(\eta^*_b(\widehat{H}^*_b) + \eta^*_b(\widehat{H})), \quad \hat{\theta}^*(\widehat{H}) = \exp(\hat{\beta}^*(\widehat{H})), \tag{7}$$

$$\hat{\beta}^*(\widehat{H}^c) = \hat{\beta}(\widehat{H}^c) - (1/B)\sum_{b=1}^{B}(\eta^*_b(\widehat{H}^{c*}_b) + \eta^*_b(\widehat{H}^c)), \quad \hat{\theta}^*(\widehat{H}^c) = \exp(\hat{\beta}^*(\widehat{H}^c)). \tag{8}$$

The bootstrap samples are drawn independently with replacement from the observed data $\{O_i := (V_i, Z_i, Y_i, \Delta_i), i = 1, \ldots, N\}$. We note that these bias corrected estimators are similar to Harrell[21] except for the inclusion of the bias terms $\eta^*_b(\widehat{H})$ and $\eta^*_b(\widehat{H}^c)$.

While variance estimates generally require double-bootstrapping, we apply the (infinitesimal) jacknife approximation[22,23], viewing (7) and (8) as "bagged estimators" (See also Rosenkranz-Ballarini[24,8] for a related context).

We describe the variance estimation for $\hat{\beta}^*(\widehat{H})$ given by (7); the variance for the complement (8) is completely analogous. Let $O^*_b = \{O^*_{b1}, O^*_{b2}, \ldots, O^*_{bN}\}$ denote bootstrap sample $b = 1, \ldots, B$ which we write as $\{O^*_{bj}, j = 1, \ldots, N\}$. Let $K^*_{bi} = \#\{O^*_{bj} = O_i\}$ be the number of times the i'th observation O_i is drawn for the b'th bootstrap sample, and let $\bar{K}^*_i = (1/B)\sum_{b=1}^{B} K^*_{bi}$.

The intinitesimal jacknife variance estimate for $\hat{\beta}^*(\widehat{H})$[22,23] is given by

$$\tilde{V} = \sum_{i=1}^{N} \widetilde{cov}_i^2, \quad \widetilde{cov}_i = (1/B)\sum_{b=1}^{B}(K^*_{bi} - \bar{K}^*_i)\big(\hat{\beta}(\widehat{H}) - \eta^*_b(\widehat{H}^*_b) - \eta^*_b(\widehat{H}) - \hat{\beta}^*(\widehat{H})\big),$$

with bias-corrected version $\hat{V}$[23] where

$$\hat{V} = \tilde{V} - \frac{N}{B}\tilde{\sigma}^2_B, \quad \tilde{\sigma}^2_B = (1/B)\sum_{b=1}^{B}\big(\hat{\beta}(\widehat{H}) - \eta^*_b(\widehat{H}^*_b) - \eta^*_b(\widehat{H}) - \hat{\beta}^*(\widehat{H})\big)^2. \tag{9}$$

In this work, the variance estimate for the bias-corrected parameter estimator will be given by $\hat{V}$ in (9) and 95% confidence intervals for hazard ratios $\hat{\theta}^*(\widehat{H})$ and $\hat{\theta}^*(\widehat{H}^c)$, in (7) and (8), will be based on standard normal approximations (exponentiated). For the FS_l and FS_{lg} estimators the lasso and GRF_{60} algorithms are mimicked for the bootstrap versions.

For summarizing estimation properties we consider bias with respect to three targets described below. Recall, the hazard function for subjects with covariate vector characteristics $\boldsymbol{Z} = \boldsymbol{z}$ with treatment set to v (0 under control, 1 under treatment) is given by $\lambda_v(t;\boldsymbol{z}) = \lambda_0(t)\exp(\gamma_0 \mathrm{v} + \gamma_1 \mathrm{v} z_1 z_3 + \boldsymbol{\gamma_2}'\boldsymbol{z}_2)$, and define $\theta_v(t) = E_{\boldsymbol{Z}}\lambda_v(t;\boldsymbol{Z})$ with the expectation over the joint covariate distribution. We define the *controlled direct effect (CDE)* of treatment as $\theta^\ddagger = \theta_1(t)/\theta_0(t)$[16], and for a generic subgroup G, $\theta^\ddagger(G)$ is defined as the above with integration restricted to G (e.g., if G is defined by $\{Z_1 = 1\} \cap \{Z_4 = 1\}$). In particular, for the true subgroups H and H^c,

$$\theta^\ddagger(H) = \exp(\gamma_0 + \gamma_1), \quad \text{and} \quad \theta^\ddagger(H^c) = \exp(\gamma_0).$$

Now, for estimated subgroups (which will generally consist of a mixture of subjects from H and H^c) we use the empirical sample version of the above expectations. That is, for subjects in $\widehat{H}$ let $\bar{\theta}_v(t;\widehat{H}) = \lambda_0(t)\exp(\gamma_0 v)\sum_{i\in\widehat{H}}\exp(\gamma_1 \mathrm{v} z_{i,1} z_{i,3} + \boldsymbol{\gamma_2}'\boldsymbol{z}_{i,2})$ and define

$$\theta^\ddagger(\widehat{H}) = \bar{\theta}_1(t;\widehat{H})/\bar{\theta}_0(t;\widehat{H}) = \exp(\gamma_0)\frac{\sum_{i\in\widehat{H}}\exp(\gamma_1 z_{i,1} z_{i,3} + \boldsymbol{\gamma_2}'\boldsymbol{z}_{i,2})}{\sum_{i\in\widehat{H}}\exp(\boldsymbol{\gamma_2}'\boldsymbol{z}_{i,2})}, \tag{10}$$

where recall for subjects in H, $z_{i,1}z_{i,3} \equiv 1$ so the above reduces to $\theta^\ddagger(H)$ if $\widehat{H} \equiv H$ ($\widehat{H}$ only consists of subjects in H). Similarly, define $\theta^\ddagger(\widehat{H}^c)$ with $\widehat{H}$ substituted with $\widehat{H}^c$ in equation (10).

Recall that for each simulated dataset the γ parameters are (known and) fixed for each model $M_1 - M_3$, however the covariates are randomly drawn from the "super-population". Therefore, even for two datasets with the same definition of $\widehat{H} \neq H$, for example $\{Z_1 = 1\} \cap \{Z_4 = 1\}$, the $\theta^\ddagger(\widehat{H})$ quantities will vary for each simulated dataset due to variation in the other covariates. The CDE's $\theta^\ddagger(\widehat{H})$ and $\theta^\ddagger(\widehat{H}^c)$ are thus random quantities with respect to $\widehat{H}$ and the covariates.

We consider bias and 95% CI coverage properties for the FS_{lg} estimator, as well as the oracle estimator (i.e., under the ideal scenario where the true H and H^c subgroups are known a-priori). Let $\hat{\theta}(H)$ denote the oracle Cox estimator, with $\hat{\theta}(\widehat{H})$ and $\hat{\theta}^*(\widehat{H})$ the observed (un-adjusted for estimation/selection of H) and bootstrap bias-corrected (via (7)) versions based on the FS_{lg} procedure, respectively. We consider three targets for each estimator ($\hat{\theta}(H)$, $\hat{\theta}(\widehat{H})$, and $\hat{\theta}^*(\widehat{H})$): the oracle estimator $\hat{\theta}(H)$, $\theta^\ddagger(\widehat{H})$, and $\theta^\dagger(H)$. For each estimator define the % relative biases: $\hat{b}^{oracle}$, $\hat{b}^\ddagger$, and $b^\dagger$ which are relative to $\hat{\theta}(H)$, $\theta^\ddagger(\widehat{H})$, and $\theta^\dagger(H)$, respectively. For example, for $\hat{\theta}(\widehat{H})$: $\hat{b}^{oracle} = (\hat{\theta}(\widehat{H}) - \hat{\theta}(H))/\hat{\theta}(H)$, $\hat{b}^\ddagger = (\hat{\theta}(\widehat{H}) - \theta^\ddagger(\widehat{H}))/\theta^\ddagger(\widehat{H})$, and $b^\dagger = (\hat{\theta}(\widehat{H}) - \theta^\dagger(H))/\theta^\dagger(H)$, which are multiplied by 100 to represent % relative bias. Define corresponding coverage measures as $\hat{C}^{oracle}$, $\hat{C}^\ddagger$, and $C^\dagger$ to indicate (for each estimator) whether the 95% confidence interval covers the respective target. For example, for $\hat{\theta}^*(\widehat{H})$, $\hat{C}^\ddagger$ is the proportion of times the 95% CI for $\hat{\theta}^*(\widehat{H})$ includes (the random) $\theta^\ddagger(\widehat{H})$.

Table 2 summarizes the properties of the FS_{lg} estimator under models $M_1 - M_3$ when additional noise factors are included (identification properties summarized in Table 1). Summaries are based on estimable (evaluable) realizations (across 1,000 simulations and $B = 300$ bootstraps) where $\{\hat{\theta}^*(\widehat{H}), \hat{\theta}^*(\widehat{H}^c)\}$ exists.

For the observed $\hat{\theta}(\widehat{H})$ the (average) relative bias, $b^\dagger$ [$\hat{b}^\ddagger$] ranges from approximately 9.2% to 24% [9.0% to 14%] across the $M_1 - M_3$ models; Indicating a general over-estimation for Cox hazard ratios based on estimated subgroups. In contrast, for the bias-corrected $\hat{\theta}^*(\widehat{H})$, $b^\dagger$ [$\hat{b}^\ddagger$] ranges from −10% to −2.4% [−11.60% to −6.3%]. For $\hat{\theta}(\widehat{H}^c)$, $b^\dagger$ [$\hat{b}^\ddagger$] ranges from 0.5% to 5.1% [−9.7% to 2.8%], and for $\hat{\theta}^*(\widehat{H}^c)$, $b^\dagger$ [$\hat{b}^\ddagger$] ranges from 2.3% to 10.9% [−4.8% to 4.6%].

For standard deviation and CI accuracy we summarize the results for $\hat{\theta}^*(\widehat{H}^c)$ under model M_3 which has the highest difference between $\theta^\dagger(H^c) = 0.56$ and $\theta^\ddagger(H^c) = 0.49$. Here the standard deviations for $\hat{\theta}(\widehat{H}^c)$ are 0.13 (for the empirical SD's) versus 0.11 (for the average of the estimated SD's) with (slight) under-coverage for $\theta^\dagger(H^c)$ ($C^\dagger = 92\%$) and under-coverage for $\theta^\ddagger(\widehat{H}^c)$ ($\hat{C}^\ddagger = 76\%$). In contrast, the standard deviations for $\hat{\theta}^*(\widehat{H}^c)$ are 0.14 versus 0.17 (resp.) with coverage rates $C^\dagger = 97\%$ and $\hat{C}^\ddagger = 93\%$.

In this setting, under models $M_1 - M_3$, the bootstrap bias-corrected estimators tend to be conservative: Under-estimating both $\theta^\dagger(H)$ and $\theta^\ddagger(\widehat{H})$ ("conservative for harm") while over-estimating both $\theta^\dagger(H^c)$ and $\theta^\ddagger(\widehat{H}^c)$ ("conservative for benefit"), except for under model M_3 where $\hat{b}^\ddagger \approx -4.8\%$. In addition, the coverage rates for $\hat{\theta}^*(\widehat{H}^c)$ are $\geq 93\%$ for each target, and the oracle coverage rates ($\hat{C}^{oracle}$) for the observed and bias-corrected estimators are $\geq 95\%$. That is, the observed and bias-corrected versions of $\widehat{H}$ and $\widehat{H}^c$ cover ($\geq 95\%$) their respective oracle counterparts.

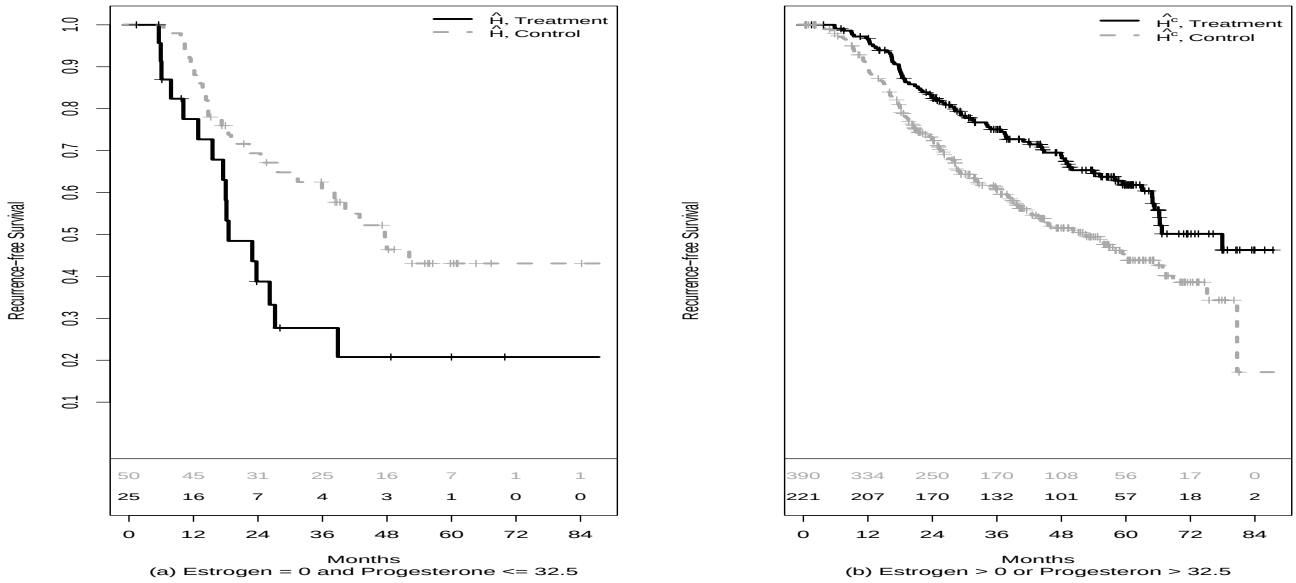

FIGURE 3 GBSG analysis application of Forest Search. Kaplan-Meier (K-M) curves (un-adjusted for the estimation of subgroups): (a) Forest Search $\widehat{H}$ subgroup treatment estimates; (b) Forest Search $\widehat{H}^c$ subgroup K-M treatment estimates.

4 | APPLICATIONS

4.1 | GBSG Analysis

In our first application we return to the GBSG trial data[10] described in Section 3. 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 (95% CI= 0.54 − 0.89). There were seven prognostic factors collected: `Estrogen`, `Age`, `Progesterone`, `Meno`, `Nodes`, `Size`, and `Grade`. The factors `Meno` and `Grade3` (`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`, `Grade3`, `Nodes`, and `Progesterone`. Applying GRF (GRF_{60} with a 6-month RMST criterion) selects `Estrogen<=0`. Now, the analyses of Sauerbrei[15] 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 binary candidate factors: `(Size<=med(Size))`; `(Nodes<=med(Nodes))`; `(Progesterone<=med(Progesterone))`; `(Estrogen<=0)`; `(Age<=40)`; `(Age<=55)`; and `Grade3`. There are then $2^{14} - 1 = 16,383$ possible all-factor combinations. However, among subgroups formed by at most 2 factors with sample sizes ≥ 60 with at least 10 events in each arm, the number of candidate subgroups is dramatically reduced to 68.

The FS_{lg} approach, maximizing consistency, estimates $\hat{H}$ as the subgroup formed by the combination of `Estrogen<=0` and `Progesterone<=32.5` (The consistency rate is 99.25%). That is, $\hat{H}$ 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 $\hat{\theta}(\widehat{H}) = 2.22$ [95% CI= 1.18, 4.2] and (bootstrap bias-corrected) $\hat{\theta}^*(\widehat{H}) = 1.53$ [95% CI= 0.78, 2.99]. For the complement, $\hat{\theta}(\widehat{H}^c) = 0.61$ [95% CI= 0.46, 0.79] and $\hat{\theta}^*(\widehat{H}^c) =$ 0.62 [95% CI= 0.42, 0.91].

The bias-corrected estimate for H^c suggests a stronger benefit (0.62 vs 0.69 for ITT) that is statistically significant and corresponds to $\approx (686 - 75)/686 = 89\%$ of the ITT population. Whereas, for H the bias-corrected estimate is not statistically significant for detriment but may suggest careful consideration for subjects without positive estrogen levels and lower levels of progresterone (≤ 32.5). Figure 3 displays the Kaplan-Meier curves for the estimated subgroups.

4.2 | ACTG-175 Analysis

In our second application we analyse subjects' survival outcomes (AIDS defining events) from the AIDS Clinical Trials Group Protocol 175 study[11]. The dataset is publicly available in the R `speff2trial` package[25].

Here our goal is to identify whether a subgroup exists with a pronounced treatment benefit. We consider the comparison of the combination treatment regimen, zidovudine and didanosine (experimental), to the monotherapy didanosine (control) treatment regimen ($N = 1083$). The Cox ITT HR estimate (with only treatment as covariate) is 0.84 (95% CI= 0.65-1.09).

For the evaluation of a large treatment benefit we switch the roles of treatment to identify a detrimental effect for control which would correspond to a potentially substantial benefit for the experimental treatment. We then simply invert the hazard ratio estimates. In addition, for the FS_{lg} consistency selection criterion we select the largest subgroup with a consistency rate of at least 90%. That is, we are searching for the largest subgroup that is "highly consistent with benefit". To this end we set the screening threshold in Step 3(a) to $\log(1/0.6)$ and the consistency threshold in Step 3(b) to $\log(1.25)$. Cox hazard ratio estimates for candidate subgroups are therefore ≤ 0.60 and random splits are ≤ 0.8 in favor of treatment. We denote the estimated subgroups by $\hat{Q}$ and $\hat{Q}^c$ (as opposed to $\hat{H}$ and $\hat{H}^c$).

Now, the survival outcomes are defined as the first occurence of three events: A decline in subjects' CD4 T cell count of at least 50; An event indicating progression to AIDS; or death. We consider the following sixteen 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), `Zprior` (prior zidovudine), `Symptom`, `Preanti` (days of prior antiviral therapy), `Str2` (0=naive antiviral history,1=experienced), and `Z30` (zidovudine 30 days prior to study).

We note that for GRF (candidate selection) we use $\tau \approx 27$ months for the truncation point and a 2 month RMST criterion (benefit of 2 months in favor of control); The 2 month threshold was chosen as a reasonable criterion since the ITT upper bound for RMST was ≈ 1.5 (in favor of control). 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[6] suggest an inflection point for age between 30 and 40 years (See their Figure 4).

With nine candidate (binary) factors there are 262, 143 all-possible combinations and the number of subgroups meeting the FS criteria was 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` (consistency rate $\approx 99\%$). That is, subjects older than 34 years (the median age) and with prior antiviral treatment for less than ≈ 1.1 years ($\approx 60\%$ of ITT) are estimated to potentially derive "consistent benefit". The resulting $\widehat{Q}$-estimates are $\hat{\theta}(\widehat{Q}) = 0.4$ [95% CI= 0.22, 0.73] and (bootstrap bias-corrected) $\hat{\theta}^*(\widehat{Q}) = 0.49$ [95% CI= 0.28, 0.87]. For the complement, $\hat{\theta}(\widehat{Q}^c) = 1.05$ [95% CI= 0.78, 1.4] and $\hat{\theta}^*(\widehat{Q}^c) = 0.97$ [95% CI= 0.69, 1.35]. The bias-corrected estimate $\hat{\theta}^*(\widehat{Q})$ suggests a relatively strong benefit (0.49 vs 0.84 for ITT) that is statistically significant and corresponds to $\approx 29\%$ of the ITT population (Figure 4 displays the Kaplan-Meier curves for the estimated subgroups).

5 | DISCUSSION

We have proposed a relatively simple and transparent approach for subgroup identification based on Cox hazard ratio estimates and a consistency criteria indicative of detrimental effects. The operating characteristics for scenarios/criteria of interest can be quickly approximated via equation (3); For example if looser/tighter control of the type-1 error is desired the screening and splitting consistency thresholds can be adjusted. In our simulations we have found the screening and splitting consistency thresholds of 1.25 and 1.0 (resp.) have good operating characteristics for identification as well as estimation. The *splitting consistency criteria* is similar in spirit to cross-validation, however in contrast to prediction, our (relatively simpler) goal is to have independent assessments for evidence of harm which is provided by both (independent) random splits having hazard ratio estimates ≥ 1.0 across repeated sample splitting. We utilize lasso and GRF for selecting candidate factors which are the basis for defining subgroups with a bootstrap bias-correction to account for the (entire) subgroup selection process.

While subgroups corresponding to the maximum (detrimental) hazard ratio estimate will generally be subgroup candidates, the proposed consistency criteria for identification does not necessarily correspond to the maximum. Recent work by Guo[7,26] considers inference for subgroups corresponding to the maximum treatment estimate. However their approach is not for general subgroup identification per se but utilized for inference when a limited set of subgroups are examined. For example in their simulations a maximum of 12 pre-specified subgroups are considered, whereas in our three applications (simulated and two real

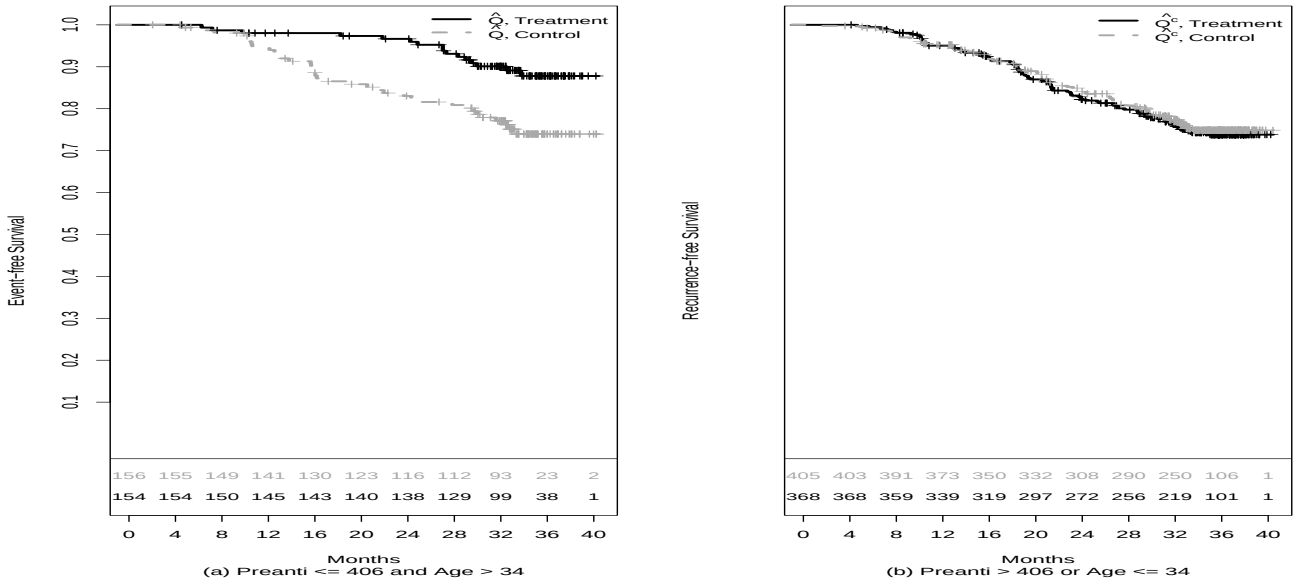

FIGURE 4 ACTG-175 analysis application of Forest Search. Kaplan-Meier (K-M) curves (un-adjusted for the estimation of subgroups): (a) Forest Search $\widehat{Q}$ subgroup treatment estimates; (b) Forest Search $\widehat{Q}^c$ subgroup K-M treatment estimates.

data) the number of subgroups examined was 70, 68, and 151, respectively. In this work we are not pre-specifying a limited set of subgroups but identifying (searching for) subgroups among a large collection of combinations, conceptually a large forest plot. In general there is not a pre-defined maximum number of subgroups.

In oncology applications the gold-standard primary analysis is the Cox model, often stratified by randomization stratification factors. To simplify we have considered the basic Cox model analysis with only the treatment arm as a covariate which is commonly used in oncology forest plot analyses. We are therefore implicitly targeting marginal hazard ratio effects, which can be quite different than the *controlled direct effect (CDE)* of treatment [16]. In the simulation study of Aalen [16] (See their Table 1) the basic Cox model was biased (over-estimated) for the ITT analysis in the presence of a single binary covariate factor which was "a highly influential risk factor" (Cox regression effect of log(4) = 1.386). In our simulations the largest discrepancy between the marginal and CDE effects, $\theta^{\dagger}(\cdot)$ vs $\theta^{\ddagger}(\cdot)$, were under model M_3 where $\theta^{\dagger}(H) = 2.0$, $\theta^{\ddagger}(H) = 2.56$, $\theta^{\dagger}(H^c) = 0.56$, and $\theta^{\ddagger}(H^c) = 0.49$. The largest covariate effect under M_3 was $\beta_5 = 0.782$ corresponding to the binary factor Z_4 in model (4). Under M_3 the % relative bias and coverage of $\hat{\theta}^*(\widehat{H})$ for $\theta^{\ddagger}(\widehat{H})$ was approximately -11.6% and 89% (See Table 2 for M_3).

While the CDEs $\theta^{\ddagger}(\cdot)$, and other estimands such as RMST are of great interest, the gold-standard analysis remains the stratified Cox model in oncology (See Freidlin [27] for a related discussion) where stratification is generally a limited set of baseline factors and not a "richly" covariate-adjusted model. We therefore consider the marginal effects of $\theta^{\dagger}(H)$ and $\theta^{\dagger}(H^c)$ to be the targets of interest which are practically obtainable.

In our simulation setting the FS and GRF_{60} approaches generally outperform the virtual twins approaches in terms of controlling the type-1 error, power, and classification accuracy. Recall that the virtual twins approaches compare survival differences at 24 and 36 months while GRF_{60} evaluates RMST over (on average) a median horizon of 48 months (range of 33-50) whereas the FS approach has no restriction and GRF compares survival differences over (on average) a median of 79 months (range of 55-84). We have considered GRF_{60} due to the fairly heavy censoring which, generated by a Weibull model based on the observed gbsg data, depends on covariates with an overall rate of approximately 46%.

The FS approach has favorable performance overall in view of the elevated type-1 errors for GRF, especially under models M_1 and M_2. When random noise factors are included in the analyses the GRF approach is more susceptible to falsely identifying subgroups. The FS_l approach is the most stable with a slight decrease in performance, while FS_{lg} inherits an increased type-1 error by the utilization of GRF_{60}, but to a much lesser extent than GRF_{60} itself. Under M_3, when there is the strongest ITT treatment effect under the null, the type-1 errors for GRF are dramatically decreased (from $\approx$ 60% to 13% under M_3.). Under the nulls of $M_1 - M_3$ the ITT treatment differences with respect to RMST are $\approx$ 7.2, 7.4, and 11.5 months. The ITT Cox marginal

effects $\theta^{\dagger}(ITT)$ are ≈ 0.7 under M_1 and M_2 and 0.55 under M_3. While $\theta^{\dagger}(ITT)$ values in the range of 0.55 is plausible, it seems more prudent to consider $\theta^{\dagger}(ITT)$'s in the range of 0.7 as more realistic in most oncology trials. Accordingly, although a limited simulation study, the FS approach may strike a more favorable balance between falsely identifying subgroups and reasonable accuracy when large subgroup effects are present. In terms of estimation, the FS_{lg} bootstrap bias-corrected estimators tend to be conservative: Under-estimating both $\theta^{\dagger}(H)$ and $\theta^{\ddagger}(\widehat{H})$ ("conservative for harm") while over-estimating both $\theta^{\dagger}(H^c)$ and $\theta^{\ddagger}(\widehat{H}^c)$ ("conservative for benefit"), except for under model M_3 where the relative bias for $\theta^{\ddagger}(\widehat{H}^c)$ is $\hat{b}^{\ddagger} \approx -4.8\%$. Though (trending) conservative, the coverage rates for $\hat{\theta}^{*}(\widehat{H}^c)$ are $\geq 93\%$ for each target, and the oracle coverage rates ($\hat{C}^{oracle}$) for $\hat{\theta}^{*}(\widehat{H})$ and $\hat{\theta}^{*}(\widehat{H}^c)$ are $\geq 95\%$. That is, the bias-corrected versions of $\widehat{H}$ and $\widehat{H}^c$ cover ($\geq 95\%$) their respective oracle counterparts.

In principle, our approach is exploratory and could be used to guide future trial development, hypothesis generation, and in conjunction with pre-specified subgroup evaluations. We believe exploratory subgroup identification is valuable even when pre-specified subgroups are of interest (e.g., biomarkers). As Zhao[28] write "A priori subgroup analyses are free of selection bias and are frequently used in clinical trials and other observational studies. They do discover some effect modification, often convincingly, from the data, but since the potential effect modifiers are determined a priori rather than using the data, many real effect modifiers may remain undetected". In practice, it is not uncommon for forest plot analyses to suggest possible concerns for an important subgroup[2]. The proposed FS approach could be utilized to provide a robust assessment that may reveal a more accurate characterization of differential treatment effects.

6 | DATA AVAILABILITY STATEMENT

The data analyzed in the applications is publicly available. R code for the simulated and real data applications, as well as for replicating the simulation results is available at the GitHub repository: https://github.com/larry-leon/ForestSearch. Additional details (e.g., computing time for analysis applications) and further simulations are also available in the repository.

References

1. *European Medicines Agency: Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials*. 2019.
2. Amatya AK, Fiero MH, Bloomquist EW, et al. Subgroup Analyses in Oncology Trials: Regulatory Considerations and Case Examples. *Clinical Cancer Research* 2021; 27(21): 5753-5756.
3. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 2011; 39(5): 1–13.
4. Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics* 2019; 47(2): 1148–1178.
5. Athey S, Wager S. Policy learning with observational data. *Econometrica* 2021; 89(1): 133–161.
6. Cui Y, Kosorok MR, Sverdrup E, Wager S, Zhu R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 2023.
7. Guo X, He X. Inference on Selected Subgroups in Clinical Trials. *Journal of the American Statistical Association* 2021; 116(535): 1498-1506.
8. Ballarini NM, Thomas M, Rosenkranz GK, Bornkamp B. subtee: An R Package for Subgroup Treatment Effect Estimation in Clinical Trials. *Journal of Statistical Software* 2021; 99(14): 1–17.
9. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011; 30(24): 2867-2880.
10. Schumacher M, Bastert G, Bojar H, et al. 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* 1994; 12(10): 2086-2093.

11. Hammer SM, Katzenstein DA, Hughes MD, et al. 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* 1996; 335(15): 1081-1090.

12. Jennison C, Turnbull BW. Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* 1984; 5(1): 33–45.

13. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2021.

14. Therneau TM. *A Package for Survival Analysis in R*. 2023. R package version 3.5-0.

15. Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M. Modelling the effects of standard prognostic factors in node-positive breast cancer. German Breast Cancer Study Group (GBSG). *British journal of cancer* 1999; 79(11-12): 1752-1760.

16. Aalen O, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect?. *Lifetime data analysis* 2015: 579–593.

17. Fan J, Gijbels I. *Local polynomial modelling and its applications*. Routledge . 2018.

18. Steingrimsson JA, Diao L, Strawderman RL. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association* 2019; 114(525): 370–383.

19. Vieille F, Foster J. *aVirtualTwins: Adaptation of Virtual Twins Method from Jared Foster*. 2018. R package version 1.0.1.

20. Tibshirani J, Athey S, Sverdrup E, Wager S. *grf: Generalized Random Forests*. 2022. R package version 2.2.1.

21. Harrell Jr. FE, Lee KL, Mark DB. MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. *Statistics in Medicine* 1996; 15(4): 361-387.

22. Efron B. Estimation and Accuracy After Model Selection. *Journal of the American Statistical Association* 2014; 109(507): 991–1007.

23. Wager S, Hastie T, Efron B. Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *Journal of Machine Learning Research* 2014; 15: 1625-1651.

24. Rosenkranz GK. Exploratory subgroup analysis in clinical trials by model selection. *Biometrical Journal* 2016; 58(5): 1217-1228.

25. Juraska M, Peter B. Gilbert w. c. f, Lu X, Zhang M, Davidian M, Tsiatis AA. *speff2trial: Semiparametric Efficient Estimation for a Two-Sample Treatment Effect*. 2022. R package version 1.0.5.

26. Guo X, Wei W, Liu M, Cai T, Wu C, Wang J. Assessing the Most Vulnerable Subgroup to Type II Diabetes Associated with Statin Usage: Evidence from Electronic Health Record Data. *Journal of the American Statistical Association* 2023; 0(0): 1-12.

27. Freidlin B, Korn EL. Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?. *Journal of Clinical Oncology* 2019; 37(35): 3455-3459.

28. Zhao Q, Small DS, Ertefaie A. Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2022; 84(2): 382-413.

TABLE 1 Average subgroup identification and classification rates across 20,000 simulations of trials under three data generation scenarios: $M_1(N = 700)$, $M_2(N = 500)$, and $M_3(N = 300)$.

	Analysis with No additional noise factors						**Analysis with additional noise factors**					
	FS_l	FS_{lg}	GRF	GRF_{60}	$VT(24)$	$VT(36)$	FS_l	FS_{lg}	GRF	GRF_{60}	$VT(24)$	$VT(36)$
						M_1 **Null:** $N = 700$, $\theta^\dagger(ITT) = 0.7$						
$any(H)^a$	0.02	0.03	**0.25**	0.05	0.03	0.04	0.02	**0.11**	**0.61**	**0.27**	0.04	0.06
$sens(\widehat{H}^C)$	1	1	0.97	0.99	1	1	1	0.99	0.92	0.97	1	0.99
$ppv(\widehat{H}^C)$	1	1	1	1	1	1	1	1	1	1	1	1
$avg\|\widehat{H}\|$	114	99	88	78	78	79	126	91	94	81	79	81
						M_1 **Alt:** $N = 700$, $p_H = 13\%$, $\theta^\dagger(H) = 2$, $\theta^\dagger(H^c) = 0.65$, $\theta^\dagger(ITT) = 0.71$						
$any(H)^b$	0.77	0.86	0.94	0.72	0.49	0.47	0.71	0.83	0.94	0.71	0.44	0.42
$sens(\widehat{H})$	0.72	0.82	0.84	0.66	0.46	0.42	0.64	0.74	0.66	0.52	0.37	0.34
$sens(\widehat{H}^C)$	0.99	0.99	0.97	0.98	0.99	0.99	0.98	0.98	0.93	0.96	0.99	0.99
$ppv(\widehat{H})$	0.69	0.8	0.78	0.61	0.44	0.41	0.6	0.71	0.6	0.47	0.36	0.33
$ppv(\widehat{H}^C)$	0.96	0.98	0.98	0.96	0.93	0.93	0.95	0.97	0.95	0.94	0.92	0.92
$avg\|\widehat{H}\|^g$	94	92	102	99	92	93	96	93	106	101	92	93
						M_2 **Null:** $N = 500$, $\theta^\dagger(ITT) = 0.69$						
$any(H)^c$	0.02	0.03	**0.23**	0.05	0.03	0.04	0.03	**0.14**	**0.6**	**0.32**	0.04	0.06
$sens(\widehat{H}^C)$	1	0.99	0.96	0.99	1	0.99	0.99	0.98	0.89	0.95	0.99	0.99
$ppv(\widehat{H}^C)$	1	1	1	1	1	1	1	1	1	1	1	1
$avg\|\widehat{H}\|$	114	100	87	76	76	80	117	88	89	80	77	79
						M_2 **Alt:** $N = 500$, $p_H = 20\%$, $\theta^\dagger(H) = 2$, $\theta^\dagger(H^c) = 0.69$, $\theta^\dagger(ITT) = 0.79$						
$any(H)^d$	0.92	0.96	0.98	0.83	0.66	0.64	0.89	0.96	0.99	0.86	0.56	0.53
$sens(\widehat{H})$	0.84	0.88	0.87	0.73	0.59	0.56	0.77	0.81	0.7	0.58	0.44	0.4
$sens(\widehat{H}^C)$	0.98	0.98	0.94	0.94	0.98	0.98	0.97	0.96	0.88	0.89	0.97	0.97
$ppv(\widehat{H})$	0.84	0.88	0.79	0.66	0.59	0.56	0.77	0.81	0.62	0.51	0.43	0.4
$ppv(\widehat{H}^C)$	0.96	0.97	0.97	0.94	0.91	0.91	0.95	0.95	0.92	0.9	0.88	0.87
$avg\|\widehat{H}\|^h$	102	101	116	115	102	103	103	101	118	119	101	102
						M_3 **Null:** $N = 300$, $\theta^\dagger(ITT) = 0.55$						
$any(H)^e$	0	0	0.05	0.01	0.01	0.02	0	0.02	**0.13**	0.07	0.01	0.02
$sens(\widehat{H}^C)$	1	1	0.99	1	1	1	1	1	0.97	0.98	1	1
$ppv(\widehat{H}^C)$	1	1	1	1	1	1	1	1	1	1	1	1
$avg\|\widehat{H}\|$	76	75	74	70	70	71	76	74	74	71	70	72
						M_3 **Alt:** $N = 300$, $p_H = 30\%$, $\theta^\dagger(H) = 2$, $\theta^\dagger(H^c) = 0.56$, $\theta^\dagger(ITT) = 0.74$						
$any(H)^f$	0.89	0.92	0.97	0.82	0.61	0.63	0.88	0.93	0.96	0.87	0.51	0.53
$sens(\widehat{H})$	0.73	0.78	0.87	0.72	0.49	0.52	0.68	0.71	0.73	0.62	0.36	0.37
$sens(\widehat{H}^C)$	0.97	0.97	0.93	0.93	0.97	0.97	0.96	0.95	0.88	0.87	0.95	0.95
$ppv(\widehat{H})$	0.8	0.84	0.83	0.68	0.53	0.55	0.76	0.78	0.7	0.59	0.39	0.4
$ppv(\widehat{H}^C)$	0.9	0.92	0.95	0.9	0.83	0.85	0.88	0.89	0.89	0.85	0.79	0.8
$avg\|\widehat{H}\|^i$	82	83	96	97	82	86	80	81	95	96	83	85

Probabilities for FS via approximation (3): [a] 0.036; [b] 0.9; [c] 0.033; [d] 0.92; [e] 0.007; [f] 0.91.
Average size of true H: [g] 89; [h] 101; [i] 90.

TABLE 2 Estimation properties for FS_{lg} under models $M_1 - M_3$ (corresponding to Table 1) across 1,000 simulations with summaries based on estimable realizations where subgroup estimates $\widehat{H}$ were obtained ($B = 300$ bootstraps): Average of the estimates (Avg); Empirical standard errors (SD); Average of estimated SD's ($\widehat{SD}$); min and max; relative biases ($\hat{b}^{oracle}$, $\hat{b}^{\ddagger}$, $b^{\dagger}$); Average CI length (Length); and Average CI coverage ($\hat{C}^{oracle}$, $\hat{C}^{\ddagger}$, $C^{\dagger}$).

	Avg	SD	$\widehat{SD}$	min	max	$\hat{b}^{oracle}$	$\hat{b}^{\ddagger}$	$b^{\dagger}$	Length	$\hat{C}^{oracle}$	$\hat{C}^{\ddagger}$	$C^{\dagger}$
M_1 $\widehat{H}$: 839 estimable realizations, Avg size of $H = 89$, $\theta^{\dagger}(H) = 2$, $\theta^{\ddagger}(H) = 2.25$												
$\hat{\theta}(H)$	2.22	0.58	0.57	1.06	6.20	0.00	-1.12	11.21	2.35	1.00	0.97	0.96
$\hat{\theta}(\widehat{H})$	2.18	0.53	0.57	1.40	6.08	-0.54	14.13	9.17	2.32	0.98	0.93	0.97
$\hat{\theta}^{*}(\widehat{H})$	1.80	0.48	0.53	1.07	4.82	-18.55	-6.28	-10.04	2.21	0.95	0.87	0.91
M_1 $\widehat{H}^c$: Avg size of $H^c = 611$, $\theta^{\dagger}(H^c) = 0.65$, $\theta^{\ddagger}(H^c) = 0.6$												
$\hat{\theta}(H^c)$	0.65	0.08	0.07	0.44	0.99	0.00	8.05	0.93	0.29	1.00	0.89	0.94
$\hat{\theta}(\widehat{H}^c)$	0.65	0.08	0.07	0.44	0.90	-0.26	2.84	0.64	0.29	1.00	0.87	0.94
$\hat{\theta}^{*}(\widehat{H}^c)$	0.66	0.08	0.11	0.45	0.92	1.41	4.55	2.33	0.43	1.00	0.96	0.99
M_2 $\widehat{H}$: 949 estimable realizations, Avg size of $H = 101$, $\theta^{\dagger}(H) = 2$, $\theta^{\ddagger}(H) = 2.61$												
$\hat{\theta}(H)$	2.34	0.60	0.57	1.10	5.75	0.00	-10.27	17.17	2.33	1.00	0.93	0.92
$\hat{\theta}(\widehat{H})$	2.39	0.58	0.61	1.41	5.75	3.09	8.99	19.40	2.48	0.99	0.92	0.93
$\hat{\theta}^{*}(\widehat{H})$	1.96	0.52	0.59	1.11	4.95	-15.95	-11.09	-2.05	2.45	0.99	0.85	0.97
M_2 $\widehat{H}^c$: Avg size of $H^c = 399$, $\theta^{\dagger}(H^c) = 0.69$, $\theta^{\ddagger}(H^c) = 0.64$												
$\hat{\theta}(H^c)$	0.69	0.10	0.10	0.43	1.01	0.00	7.52	0.04	0.38	1.00	0.92	0.95
$\hat{\theta}(\widehat{H}^c)$	0.69	0.10	0.10	0.43	1.05	0.47	-1.82	0.50	0.38	1.00	0.83	0.94
$\hat{\theta}^{*}(\widehat{H}^c)$	0.71	0.11	0.14	0.44	1.12	3.49	1.12	3.56	0.56	1.00	0.94	0.98
M_3 $\widehat{H}$: 924 estimable realizations, Avg size of $H = 90$, $\theta^{\dagger}(H) = 2$, $\theta^{\ddagger}(H) = 2.56$												
$\hat{\theta}(H)$	2.29	0.61	0.60	1.00	6.97	0.00	-10.64	14.34	2.47	1.00	0.94	0.95
$\hat{\theta}(\widehat{H})$	2.48	0.62	0.73	1.45	6.97	10.21	12.62	23.97	3.04	0.99	0.95	0.95
$\hat{\theta}^{*}(\widehat{H})$	1.95	0.52	0.69	1.11	5.83	-13.66	-11.58	-2.39	2.96	1.00	0.89	0.97
M_3 $\widehat{H}^c$: Avg size of $H^c = 210$, $\theta^{\dagger}(H^c) = 0.56$, $\theta^{\ddagger}(H^c) = 0.49$												
$\hat{\theta}(H^c)$	0.55	0.11	0.11	0.25	1.10	0.00	11.31	-1.32	0.45	1.00	0.92	0.94
$\hat{\theta}(\widehat{H}^c)$	0.59	0.13	0.11	0.28	1.35	6.79	-9.69	5.14	0.45	0.99	0.76	0.92
$\hat{\theta}^{*}(\widehat{H}^c)$	0.62	0.14	0.17	0.28	1.41	12.62	-4.76	10.93	0.68	1.00	0.93	0.97