1 | RESPONSE TO REVIEWER 1

Thank you for your questions and comments. References to the revised text are in italics.

1.1 | Question 1

The authors use lasso and grf for pre-selection of candidate factors. Can the authors be more specific as the outcome is survival time? Do you mean that you used regularized Cox regression and causal survival forest.

Thank you for your comment. We have clarified in the 3rd paragraph of the introduction which now reads:

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 candidate factors. For the selection of candidate factors any well-defined algorithm can be applied. As a "base-case" algorithm we consider generalized random forests,[1,2,3] henceforth GRF, as a core component which we use with or without lasso.[4] GRF is a subgroup identification approach itself based on restricted mean survival time summaries via causal survival forests, whereas lasso estimates the Cox model via regularization. In our applications we illustrate various combinations of GRF and lasso, including evaluating all baseline factors with continuous covariates cut at the quartiles. For identified subgroups, $\widehat{H}$ and $\widehat{H}^c$ say, inference based on bootstrap bias-corrected estimators is described which accounts for the overall FS algorithm including the manner in which candidate factors are selected.

Please also note that in the simulations section (Section 3) in the fourth paragraph below Equation (6) we cite the software:

Now, in addition to the proposed FS approach we evaluate virtual twins[5] and GRF[1,2,3] procedures for subgroup identification. To account for censoring with the virtual twins approach we employ a basic "censoring unbiased transformation"[6] (Doubly-robust versions are also available[7]). Virtual twins is implemented via the R package `aVirtualTwins`,[8] and generalized random forests is implemented using the `causal_survival_forest` function in the R `grf` package.[9,10] When utilized in the FS algorithm lasso is implemented with the `glmnet` R package.[4]

1.2 | Question 2

Can the authors define the type 1 error and power rigorously?

Thank you for this suggestion. We have defined the type-1 error and power in the 2nd paragraph of Section 2, below Equation (1):

In this work we assume heterogeneous treatment effects are induced by the existence of a detrimental subgroup H with true marginal hazard ratio $\theta^{\dagger}(H) > 1$ where the size of H is at least 60 subjects with an underlying expected event rate d. In our context there are two type-1 error scenarios for false subgroup identification: (i) If a subgroup H is identified where in truth $\theta^{\dagger}(H) \leq 1$ (non-detrimental); and (ii) If the treatment effect is uniformly beneficial, $\theta^{\dagger}(\text{ITT}) < 1$. Under scenario (i) it is possible for heterogeneous treatment effects to exists, but the composition of the identified subgroup H is such that treatment is non-detrimental for the sub-population ($\theta^{\dagger}(H) \leq 1$); in contrast under (ii) there does not exist such subgroup effects. In the following section (Section 2.1) we represent hazard ratio estimators based on a subgroup, and random splits thereof, via two normal random variables where the joint probability of meeting the screening and splitting consistency criterion thresholds is calculated by numerical integration. Specifically, let W_1 and W_2 be two (independent) $N(\log(\theta^{\dagger}(H)), 8/d)$ random variables and define $p(c_1, c_2; d, \theta^{\dagger}(H)) = \Pr(W_1 + W_2 \geq 2\log(c_1), \min(W_1, W_2) \geq \log(c_2))$ where c_1 and c_2 are the screening and consistency thresholds. Here W_1 and W_2 represent the Cox estimators corresponding to the random (50/50) subgroup splits, and the sum $W_1 + W_2$ represents the Cox estimator for the subgroup. For fixed d and thresholds $\{c_1, c_2\}$ the type-1 error is approximately $p(c_1, c_2; d, 0)$ for $\theta^{\dagger}(H) = 1$, and power $p(c_1, c_2; d, \theta^{\dagger}(H))$ for $\theta^{\dagger}(H) > 1$. The practical ramifications for false identification depends on the true $\theta^{\dagger}(H)$. For example, if the true treatment effect is uniform with an ITT benefit of 0.75 (which may be considered "clinically significant" in various oncology settings) then for subgroup size $n = 60$ with a censoring rate of

45% the type-1 error is approximately 4.9% for $c_1 = 1.25$ and $c_2 = 1.0$ under $\theta^\dagger(H) \approx \theta^\dagger(\text{ITT}) = 0.75$ (details are discussed in Section 2.1.

1.3 | Question 3

In the algorithm, the choice of the log hazard ratio log(1.25), log(1.0) seems pretty heuristic. Some discussions are helpful

In Section 2.1 we describe how the power is approximated by $p(c_1, c_2; d, \theta^\dagger(H))$ above which can be calculated by numerical integration which reads as follows:

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

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(w_1 + w_2 \geq 2\log(1.25))I(w_1 \geq 0)I(w_2 \geq 0)\varphi(w_1; \beta, 8/d)\varphi(w_2; \beta, 8/d)dw_1dw_2, \tag{3}$$

where $\{W_1, W_2\} \sim N(\beta, 8/d)$ (independently), and $\varphi(\cdot; \beta, 8/d)$ denotes the normal density with mean β and variance $8/d$. In Supplementary materials S1 we provide a simulation evaluation of the approximations where we find the approximations to appear quite accurate.

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 $\approx$ 10% when the underlying hazard ratio for H is below 1.0. If the underlying treatment effect is uniform and beneficial then for a random subgroup H, Cox model estimates will randomly fluctuate around the ITT effect. For example, for $\theta^\dagger(H) \equiv \theta^\dagger(\text{ITT}) = 0.75$, the above approximation is 0.049, 0.033, and 0.022 (for $n = 60$, 80, and 100, respectively) indicating reasonable control of type-1 error. We note that because FS seeks subgroups with evidence for harm (viz-a-viz the screening and consistency thresholds) the chance of forming subgroups under the null with an estimated benefit randomly in favor of control is less likely the stronger the (uniform) ITT treatment effect.

The 2nd paragraph of the discussion section also reads:

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.

1.4 | Question 4

eq (5) has text which looks weird

Sorry, we have tried to clarify the notation. Referring to the data-generating model the text is revised as follows:

The true model was

$$\log(T) = \mu + \beta_0\text{V} + \beta_1\text{V}Z_1Z_3 + \beta_2Z_1 + \beta_3Z_2 + \beta_4Z_3 + \beta_5Z_4 + \beta_6Z_5 + \tau\epsilon, \tag{4}$$

with V denoting treatment, ϵ was from the standard extreme value distribution and τ was a dispersion parameter.

Writing the above data-generating model as

$$\log(T) = \mu + \beta_0\text{V} + \beta_1\text{V}Z_1Z_3 + \boldsymbol{\beta}_2{}'\boldsymbol{Z}_2 + \tau\epsilon, \tag{5}$$

with $\boldsymbol{Z}_2 := (Z_1, Z_2, Z_3, Z_4, Z_5)$ and $\boldsymbol{\beta}_2 = (\beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ defined accordingly, 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 $\gamma = -\tau\beta$,[11] say. The parameters μ, $\boldsymbol{\beta}_2$, and τ were 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 $\theta^\dagger(H^c) = 0.65$).

1.5 | Question 5

The authors consider the Cox model as a running model for analysis. Can the method be extended to some machine learning methods such as survival forests

It is not clear how to extend the approach to machine learning methods. Our goal was to develop an approach based on standard Cox models, where *in theory*, one could conduct a Cox model analysis as if the subgroup was pre-specified (the oracle estimator) using the gold-standard (in oncology clinical settings) analysis. On the other hand, we are directly using results of the causal survival forests approach of GRF which targets RMST. That is, we consider the subgroups found by GRF as candidates and evaluate them within the framework of our Cox model criteria. This could be extended to include other machine learning methods for censored data in the same manner so that "a collection of machine learning" approaches could be evaluated with estimation in terms of the Cox model (viz-a-viz our bootstrap bias-correction and variance estimation procedure).

The first paragraph of the discussion section now reads:

We have proposed a relatively simple and transparent approach for subgroup identification based on Cox hazard ratio estimation criteria indicative of detrimental effects. We utilize optional GRF and lasso procedures for selecting candidate factors (binary splits) which are the basis for defining subgroups. GRF is itself a subgroup identification procedure which targets RMST, whereas our use of lasso is for Cox model covariate (prognostic) selection. In general, any well-defined algorithm can be implemented such as (pre-defined) clinical, and health technology assessment[12] considerations and/or various machine learning algorithms for censored data. In applications, choices for components of the FS algorithm may be better suited than others. In particular, in our simulations we found the lasso to help mitigate false-discovery when analyses include baseline factors that are completely random noise. However such random noise seems extreme in clinical trials when baseline factors generally have some degree of prognostic value; nevertheless, the lasso may aid in algorithmic stability. In addition, whether to maximize the consistency rate or to choose the largest subgroup with a high consistency rate (e.g., at least 90%) may differ in stability. While the proposed CV evaluation cannot establish optimality of a chosen algorithm, it can discern between the quality and stability of algorithms. The bootstrap bias-correction and variance estimation procedure for the resulting FS Cox hazard ratios would incorporate the chosen algorithm; however if several FS algorithms are evaluated this (exploratory) iterative aspect would not be taken into account.

The third paragraph of the discussion section describes an extension to adjusted Cox models:

Our main application is subgroup analyses for survival outcomes. In oncology applications the gold-standard primary analysis is the Cox model[13] usually 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; and is the "most common approach to analysis".[11] However, adjusted Cox models[14] can also be used either by stratification, direct covariate adjustment (with care to account for any subgroup redundancies in the model) or propensity score-weighting.[15] In addition to bootstrap bias-corrected Cox model estimates, other summaries can also be provided such as RMST and Kaplan-Meier survival curves (e.g., across pre-defined timepoints). For these summaries the bias-correction and variance estimation procedure described for the Cox model hazard ratios can be applied in an analogous manner.

And the last paragraph of the discussion mentions areas of future work:

Future work will be to extend the inclusion of GRF to additional machine learning methods for censored data so that "a collection of machine learning" approaches for FS candidate selection could be evaluated; where regardless of the (algorithmic) *source of candidate selection, the resulting FS estimation is in terms of the proposed bootstrap bias-corrected Cox model. While our bootstrap bias-correction and variance estimation appears to work well, it would be interesting to evaluate the applicability of the Guo et al*[16] *bootstrap calibration procedure to our setting which generally involves a large collection of subgroup candidates.*

2 | RESPONSE TO REVIEWER 2

Thank you for your questions and comments. Please find our responses below. References to the revised text are in italics.

2.1 | Question 1

The authors describe a procedure (based on LASSO and GRF) for selecting a set of factors to consider as candidates for the subgroup identification component of the proposed method. This procedure appears to work relatively well for the intended purpose, but seems somewhat ad hoc. I think it would be helpful if the authors gave a more detailed discussion of the intuition behind this procedure. Were there other versions of this factor selection procedure that were considered prior to landing on the current revision?

We initially only considered the combination of GRF with lasso for its simplicity, however based on your additional questions we further implemented cross-validation approaches to judge the stability and quality of the proposed algorithm (to be discussed below). We found in the real data applications that the original GRF/Lasso approach did not appear to be the best option in comparison to a broader approach for selecting candidates.

We have expanded the "base-case" algorithm to include evaluating continuous covariates with factors defined by cuts at the quartiles, with or without lasso. The 3rd paragraph of the introduction reads:

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 candidate factors. For the selection of candidate factors any well-defined algorithm can be applied. As a "base-case" algorithm we consider generalized random forests,[1,2,3] *henceforth GRF, as a core component which we use with or without lasso.*[4] *GRF is a subgroup identification approach itself based on restricted mean survival time summaries via causal survival forests, whereas lasso estimates the Cox model via regularization. In our applications we illustrate various combinations of GRF and lasso, including evaluating all baseline factors with continuous covariates cut at the quartiles. For identified subgroups, $\widehat{H}$ and $\widehat{H}^c$ say, inference based on bootstrap bias-corrected estimators is described which accounts for the overall FS algorithm including the manner in which candidate factors are selected. 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. To evaluate the quality and stability of the FS algorithm(s) we propose two forms of cross-validation.*

The second to last paragraph of the introduction reads:

The manner of choosing candidate factors (binary splits) is not restricted to the above GRF and lasso algorithm. In our applications we also consider GRF along with cutting all continuous factors at the mean, median, 1st quartile (q_1), and 3rd quartile (q_3), where lasso is not included in the algorithm (i.e., four splits for each continuous factor). For example, with 6 binary and 4 continuous factors there would be $L = 44$ subgroup indicators (22 binary factors) and 990 possible two-factor subgroup combinations. In addition, one can first apply lasso and then cut all (lasso selected) continuous factors in the above manner. Whichever candidate selection algorithm is employed the bootstrap process for bias-correction and variance estimation would incorporate the algorithm, mimicking the entire procedure. To evaluate the quality and stability of the chosen algorithm, and to compare algorithms (e.g., with or without lasso), we propose two forms of cross-validation.

In our applications we illustrate the comparison of candidate selection algorithms as well as the FS subgroup selection criterion in terms of cross-validation performance. The cross-validation approach is provided in response to the next comment below. Revised text of Section 4.1 (GBSG analysis) reads:

... There were $p = 7$ prognostic factors collected: `Estrogen`, `Age`, `Prog`, `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.*

In this analysis we select the largest subgroup with a consistency rate of at least 90% where lasso[4] *is first applied with the aforementioned factors, and for the continuous factors (selected per lasso), these are cut at the mean, median, 1st quartile, and 3rd quartile. We note that an alternative analysis where lasso is not applied yields the same estimated subgroups and virtually identical bootstrap bias-corrected estimates (described below). In addition, another alternative analysis maximizing the consistency rate is described in Supplementary materials S2.1. In comparison to these alternative analyses, the 10-fold CV properties for the current analysis suggests preferable algorithmic stability (details described below).*

Now, the first stage of our algorithm is to apply lasso which selects `Grade3`, `Size`, `Nodes`, *and* `Prog`, *the last three of which are continuous, and binary cuts at the mean, median, 1st quartile, and 3rd quartile were included for each continuous factor. Next, applying GRF (GRF_{60} with a 6-month RMST criterion) selects* `Estrogen<=0` *(Estrogen cut at 0). There were then $K = 14$ candidate factors (binary cuts) and thus $L = 28$ total single factor subgroups with $L(L-1)/2 + L = 406$ possible subgroups (two-factor combinations); among these subgroups the number of candidates with sample sizes ≥ 60 and at least 10 events in each arm was reduced to 263.*

The FS approach estimates $\widehat{H}$ as the subgroup `Estrogen<=0` *(The consistency rate is 95.1%). That is, $\widehat{H}$ subjects are those with an estrogen level of 0 and the resulting $\widehat{H}$-estimates were $\hat{\theta}(\widehat{H}) = 1.95$ (1.05, 3.61) with bootstrap bias-corrected $\hat{\theta}^*(\widehat{H}) = 1.58$ (0.86, 2.9). For the complement, $\hat{\theta}(\widehat{H}^c) = 0.61$ (0.47, 0.8) and $\hat{\theta}^*(\widehat{H}^c) = 0.64$ (0.44, 0.93) ...*

... We note that the GRF approach itself also identified `Estrogen<=0`. *In addition, when the current FS selection criteria is modified by not including lasso in the algorithm, the FS approach also identified* `Estrogen<=0` *with virtually identical $\hat{\theta}^*(\widehat{H})$ and $\hat{\theta}^*(\widehat{H}^c)$ estimates. However the 10-fold CV properties do not compare favorably to the current FS analysis. Specifically, across the 10-fold CV analyses a subgroup was identified (median) 8 out of 10 times with a sensitivity of $sensCV(\widehat{H}) = 55\%$. The positive predictive value was $ppvCV(\widehat{H}) \approx 67\%$, and for the complement, the medians for the corresponding $sensCV$ and $ppvCV$ were 96% and 94%, respectively (The exact $\widehat{H}$ subgroup definition of* `Estrogen<=0` *was reproduced (median) 40% of the time.). Moreover an additional FS analysis, maximizing consistency, estimates $\widehat{H}$ as the subgroup formed by the combination of* `Estrogen<=0` *and* `Prog<=32.5`, *which in comparison to the aforementioned FS algorithms exhibits less favorable CV properties (details in Supplementary materials S2.1).*

Similar considerations are now included in the revised ACTG-175 analysis of Section 4.2, as well as supplementary analyses (Section 4.3) provided in the Supplementary materials. In the supplementary analyses we “pressure test” the stability of the FS procedure by including 20^+ random (standard normal) "noise" covariates and find utility with the use of lasso.

We note that any well-defined (pre-specified) algorithm can be used for candidate selection. The first paragraph of the discussion section reads:

We have proposed a relatively simple and transparent approach for subgroup identification based on Cox hazard ratio estimation criteria indicative of detrimental effects. We utilize optional GRF and lasso procedures for selecting candidate factors (binary splits) which are the basis for defining subgroups. GRF is itself a subgroup identification procedure which targets RMST, whereas our use of lasso is for Cox model covariate (prognostic) selection. In general, any well-defined algorithm can be implemented such as (pre-defined) clinical, and health technology assessment[12] *considerations and/or various machine learning algorithms for censored data. In applications, choices for components of the FS algorithm may be better suited than others. In particular, in our simulations we found the lasso to help mitigate false-discovery when analyses include baseline factors that are completely random noise. However such random noise seems extreme in clinical trials when baseline factors generally have some degree of prognostic value; nevertheless, the lasso may aid in algorithmic stability. In addition, whether to maximize the consistency rate or to choose the largest subgroup with a high consistency rate (e.g., at least 90%) may differ in stability. While the proposed CV evaluation cannot establish optimality of a chosen algorithm, it can discern between the quality and stability of algorithms. The bootstrap bias-correction and variance estimation procedure for the resulting FS Cox hazard ratios would incorporate the chosen algorithm; however if several FS algorithms are evaluated this (exploratory) iterative aspect would not be taken into account.*

And the last paragraph of the discussion mentions areas of future work:

Future work will be to extend the inclusion of GRF to additional machine learning methods for censored data so that "a collection of machine learning" approaches for FS candidate selection could be evaluated; where regardless of the (algorithmic) source of candidate selection, the resulting FS estimation is in terms of the proposed bootstrap bias-corrected Cox model. While our bootstrap bias-correction and variance estimation appears to work well, it would be interesting to evaluate the applicability of the Guo et al[16] bootstrap calibration procedure to our setting which generally involves a large collection of subgroup candidates.

2.2 | Question 2

Related to the previous comment, I think it would be helpful if the authors included some discussion of the computational complexity of the proposed procedure. For example, could the authors give a sense of how large K can be before the proposed method becomes infeasible? If K can feasibly be very large, would it be worth considering a larger set of candidate dummy indicators in the subgroup identification portion of their procedure

Thank you for your comment. As mentioned in the response to the previous comment we have expanded our "base-case" algorithm to include cutting all continuous covariates at the mean, median, 1st quartile (q_1), and 3rd quartile (q_3). In addition, we have proposed two forms of cross-validation which are more computationally intensive than the FS subgroup identification analysis and bootstrapping. In the main text, the most computationally intensive analysis was for the ACTG-175 trial which included $K = 33$ candidate (binary) factors, $L = 66$ total single factor subgroups, and $L(L-1)/2 + L = 2,211$ possible subgroups formed by two factor combinations. However, in the Supplementary materials we provide additional analyses for much larger K's.

In particular, in Supplementary materials S2.3 we artificially add 20 random continuous noise factors to the ACTG-175 dataset which resulted in $K = 93$ candidate factors, $L = 186$ total single factor subgroups and consequently $17,391$ possible subgroups. The timing for the subgroup identification analysis was about 1 minute, and 54 minutes for the 2000 bootstraps. The most intensive analysis was the 200 10-fold cross-validation analysis which took around 3.7 hours but practically could be reduced to approximately half with 100 random folds.

Below we summarize the timings for the analyses. We first describe the proposed cross-validation approaches.

The first paragraphs of (introduction to) the applications section (Section 4) now reads:

In applications, as suggested by a reviewer, we consider cross-validation (CV) for evaluating the quality and stability of the selection algorithms (See also Athey and Wager[2], and Knaus[17]). Two forms of CV are implemented, 10-fold CV, and what we refer to as N-fold CV defined as follows. For N-fold CV we exclude each subject ($i = 1, \dots, N$) from the analysis and predict their $\widehat{H}$ ($\widehat{H}^c$) classification (based on the remaining $N-1$ subjects) where if a subgroup $\widehat{H}$ is not identified then the subject is classified as $\widehat{H}^c$ (i.e., $\widehat{H} = \emptyset$). That is, let $\hat{\pi}^{-i}(\boldsymbol{Z}_i)$ denote the ith subjects' predicted classification based on the FS procedure ($\widehat{H}$ or $\widehat{H}^c$) without the subject in the analysis. Similarly define $\hat{\pi}(\boldsymbol{Z}_i)$ as the FS classification based on the full sample analysis and form $\widehat{O}_{CV} = \{\hat{O}_i := (V_i, Y_i, \Delta_i, \hat{\pi}(\boldsymbol{Z}_i), \hat{\pi}^{-i}(\boldsymbol{Z}_i)), i = 1, \dots, N\}$. Cox model analyses based on $\hat{\pi}(\cdot)$ subgroups correspond to estimates that are un-adjusted for the selection algorithm whereas $\hat{\pi}^{-i}(\cdot)$ represents an out-of-bag (OOB) classification where each subject is not included in the selection algorithm from which they are classified. Correspondence between $\hat{\pi}(\cdot)$ and $\hat{\pi}^{-i}(\cdot)$ subgroup analysis results may be anticipated, especially for large N. Of course if $\hat{\pi}$ and $\hat{\pi}^{-i}$ are identical then there is no diagnostic value; in contrast substantial lack of correspondence may suggest an underlying instability. In 10-fold CV we randomly partition the data into 10 folds and for each fold (leaving these subjects out) select $\widehat{H}$ based on the other 9 folds to predict the classification for that left out fold. This yields an alternative version of $\widehat{O}_{CV}$ where $\hat{\pi}^{-i}$ now corresponds to the predicted classification based on the left out fold analysis to which the ith subject belongs. Since this process generally depends on the random partition we repeat this 200 *times and summarize correspondence measures across the partitions.*

For both CV approaches we consider metrics such as how often a subgroup is identified based on the "training samples" and the correspondence with the full sample analysis $\widehat{H}$ definitions, as well as in terms of sensitivity and positive predictive value measures. To this end, the sensitivity and positive predictive value metrics in Section 3 are modified by replacing $\widehat{H}$ with $\widehat{H}^{-i}$

(i.e., $\hat{\pi}^{-i}$) and the true H with $\widehat{H}$. For example, $sensCV(\widehat{H}) := \#\{i \in \widehat{H}^{-i} \cap \widehat{H}\}/\#\{i \in \widehat{H}\}$ denotes the correspondence between the CV "testing prediction" and the full analysis $\widehat{H}$-classification (relative to the size of the full analysis $\widehat{H}$).

In applications, choices for components of the FS algorithm may be better suited than others. In particular, in our simulations we found the lasso to help mitigate false-discovery when analyses include baseline factors that are completely random noise. However such random noise seems extreme in clinical trials when baseline factors generally have some degree of prognostic value; nevertheless, the lasso may aid in algorithmic stability. In addition, whether to maximize the consistency rate or to choose the largest subgroup with a high consistency rate (e.g., at least 90%*) may differ in stability. While the proposed CV evaluation cannot establish optimality of a chosen algorithm, it can discern between the quality and stability of algorithms.*

In our applications the 10-fold CV was the most computationally intensive in terms of timing. The timings for the analyses in the main text are provided:

The last paragraph of Section 4.1 (GBSG) analysis reads:

The computational timing for the current analysis on an Apple studio (M1 20 core with 69 GB) was approximately: 0.05 *minutes for the FS analysis;* 29 *minutes for the* 2000 *bootstraps;* 4 *minutes for the N-fold CV; and* 59 *minutes for the* 200 *random 10-fold CV analyses. In total, the number of minutes was* ≈ 92*.*

The 2nd to last paragraph of Section 4.2 (ACTG-175) analysis reads:

The computational timing for the current analysis on an Apple studio (M1 20 core with 69 GB) was approximately: 0.2 *minutes for the FS analysis;* 30 *minutes for the* 2000 *bootstraps;* 22 *minutes for the N-fold cross-validation; and* 105 *minutes for the* 200 *random 10-fold cross-validation analyses. In total, the number of minutes was* ≈ 157*.*

Analyses with larger K's are provided in the Supplementary materials. For the ACTG-175 analysis these are overviewed in the last paragraph of Section 4.2 which reads:

In Supplementary materials S2.2-S2.4 we provide additional analyses. In particular, we artificially add 20 *standard normal baseline factors as random noise candidates where the FS algorithm does not include lasso in S2.3, while in S2.4 lasso is included. When lasso is not included FS identifies a nonsensical subgroup based on a random noise factor. In contrast, when lasso is included the same subgroup as the full analysis above and (essentially) the same bootstrap bias-adjusted estimates were obtained. However, N-fold CV discrepancies suggest an underlying instability in the presence of including the* 20 *random noise factors.*

In Supplementary materials S2.3 where we now artificially add 20 random continuous noise factors ($N(0, 1)$) which resulted in $K = 93$ candidate factors, $L = 186$ total single factor subgroups and consequently $17, 391$ possible subgroups:

The timing was (Apple M1 20 core with 69 GB) 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*.*

Additional analyses with larger K's are provided in the Supplementary materials. Section 4.3 (Supplementary analyses) reads:

As described above, additional analyses of the GBSG and ACTG-175 trials are available in the Supplementary materials S2.1-S2.4. Supplementary materials S2.5 also provides analysis of the systolic heart failure data[18] *available in the* `randomForestSRC`[19] *package (a larger trial with $N = 2,231$ subjects, $p = 38$ baseline covariates, and $K = 78$); in addition, we 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 our code implements parallel computing via the* `doFuture`[20] *package for the bootstrapping and CV procedures; accordingly the timing of computations depends on the number of available cores.*

In summary we suspect that most clinical applications will have less than 100 candidate factors which we have found to be computationally feasible for the sample sizes considered in our analysis examples ($N \approx 1,000$ and $2,000$).

2.3 | Question 3

To help better illustrate how the proposed methods might be used in practice, I think it would be very helpful if the authors provided a more detailed discussion/interpretation of their findings for the two real data examples. For example, regarding the adjuvant endocrine therapy trial, the authors note that the identified subgroup is those with estrogen receptor values of 0 and PR values of 32.5 or less. Could the authors perhaps include some discussion of why this subgroup might not benefit from the additional of adjuvant hormone therapy? What is the clinical interpretation of this? (I do think it's intuitive that patients with lower levels of expression of HR and/or PR may not respond as well as to endocrine therapy, but it would be good to add some discussion of this for readers who are not familiar with breast cancer). It would also be helpful to include a similar discussion of the interpretation of the identified subgroup for the ACTG-175 trial.

Thank you for your suggestion. As discussed in the previous response, we have implemented a broader approach for the selection of candidate factors and now apply cross-validation for evaluating the stability/quality of the overall FS algorithm. The original analysis of the breast cancer study (where the subgroup consisting of subjects with estrogen receptor values of 0 and PR values of 32.5 or less are identified) is moved to the Supplementary materials S2.1. The revised analysis identifies estrogen values of 0 as the subgroup; the algorithm for which we have found to exhibit more favorable cross-validation properties compared to the FS algorithm that identified the original subgroup. The original ACTG-175 analysis was also revised. The original ACTG-175 analysis has been moved to Supplementary materials S2.2. The revised analysis is based on the same factors, Preanti (days of prior antiretroviral therapy) and Age, but the Preanti cut has been revised from `Preanti<=406` to `Preanti<=744.5`.

We have now added the following interpretation and attempted evaluation of the plausibility for the breast cancer study subgroup finding. The last paragraph of Section 4.1 containing the GBSG analysis reads:

Regarding the plausibility of the subgroup analysis results suggesting subjects without positive estrogen levels may not benefit from tamoxifen treatment compared to chemotherapy. We note that tamoxifen is a selective estrogen receptor (ER) modulator and is mainly indicated (as of the 2016 era) for the treatment of breast cancer in postmenopausal women and postsurgery neoadjuvant therapy in ER-positive breast cancers.[21] *ER-negative tumors are characterized by the lack of (or very small levels of) ER expression*[21] *with 2010 guidelines suggesting tumors with* $\geq 1\%$ *expression of ER to be considered ER positive.*[22] *In a meta-analysis of five randomized prevention trials, Cuzick et al*[23] *report that, in the tamoxifen prevention trials, there was no effect for breast cancers that were negative for estrogen receptor (hazard ratio 1.22 [0.89-1.67]). More recently, in a patient-level meta-analysis of randomized trials conducted by the Early Breast Cancer Trialists' Collaborative Group, for 'estrogen negative' (ER=0) subjects, there were over 5,000 woman-years of follow-up in each of the tamoxifen and control arms with similar events (162 events for tamoxifen and 163 for control) corresponding to an estimated event-rate ratio of 1.11 (SE=0.13); whereas for 'estrogen positive' (ER* $\geq$ *10%) subjects the event-rate ratio was 0.62 (SE=0.03).*[24] *Lastly, in Supplementary materials S3 we applied the identified subgroup definitions (ER=0, ER>0, say) to the Rotterdam tumor bank data*[25] *which was utilized for "external validation of a Cox prognostic model".*[26] *Briefly, the Rotterdam data was observational and we implemented (stabilized) propensity-score weighting.*[15] *The Cox model estimates were* $0.55\ (0.30, 1.01)$ *for subjects without estrogen levels (ER=0) and* $0.65\ (0.49, 0.86)$ *for subjects with positive estrogen levels (ER>0). In contrast to our results, estimates for subjects without estrogen levels trended towards a favorable benefit; whereas estimates for subjects with positive estrogen levels were fairly consistent compared to* $\hat{\theta}^*(\widehat{H}^c) = 0.64\ (0.44, 0.93)$.

The last paragraph of Section 4.2 containing the ACTG-175 analysis reads:

Evaluating the plausibility of the subgroups is hampered by direct access to summaries by the combination of prior antiretroviral therapy duration (`Preanti`*) and age in the literature. However, the role of prior antiretroviral therapy duration viz-a-viz naive-vs-experienced is an important aspect and frequently a key component of regimens studied: Katzenstein et al*[27] *write "Based on studies of HIV RNA suppression and the development of drug resistance, the goals of antiretroviral treatment in HIV* *infection have rapidly shifted to early suppression of HIV replication to the lowest possible levels with combination antiretroviral therapy regimens." The HIV Trialists' Collaborative Group (HIVTCG) conducted a patient-level meta-analysis (including the ACTG-175 study) of randomized trials:*[28] *In reference to the combination of zidovudine and didanosine, "there appeared to be greater effects on the rate ratio for death and disease progression among participants who, at baseline, had either no previous antiretroviral therapy or higher CD4-cell counts." Now, for the* $\widehat{Q}$ *subgroup (*`Preanti<=744.5` *and* `Age>34`*) consisting of* $n = 382$ *subjects there were* 46.9% *who were antiretroviral treatment naive; for whom the (estimated) enhanced benefit may be plausible in view of the aforementioned HIVTCG analysis. For the remaining* 53.1% *of subjects, of whom, approximately* 91% *had zidovudine use in the 30 days prior to treatment initiation and a mean prior antiretroviral therapy duration of* 352 *days at baseline (mean [median] baseline CD4 count of* 324 [320]*, min =* 70*, max =* 702*). Consequently, for these subjects, initiation of the combination of zidovudine and didanosine generally amounted to the continuation of zidovudine while adding didanosine after (on average) less than a year of prior antiretroviral therapy. We conjecture the recent zidovudine exposure/experience (as well as "moderate prior antiretroviral therapy duration") with the addition of didanosine may have helped with (subsequent) tolerability of the combination and with resistance; however this is just conjecture as we are not aware of available subgroup summaries that are directly applicable.*

2.4 | Question 4

Related to the previous comment, the authors note that the proposed methods should be applied in an exploratory fashion, which I agree seems most appropriate. However, I still think it would be helpful if the authors included a more detailed discussion of what the next steps would be (after implementing the proposed method) in practice. For example, how would they propose that their findings be used moving forward? Would the next step be validation of the findings? And how would they suggest the results be incorporated into clinical/research practice once validated?

Thank you for your comments. In addition to the plausibility evaluation for the analysis application findings that was discussed in the previous comment, we have added the following in the last few paragraphs of the Discussion section:

In principle our approach is exploratory and could be used to guide future trial development. We believe exploratory subgroup identification is valuable even when pre-specified subgroups are of interest (e.g., biomarkers). As Zhao *et al*[29] *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 our data analysis applications we consider available data sources to evaluate the plausibility of our subgroup findings. Patient-level meta-analyses of randomized trials*[24,28] *seems the most feasible and robust avenue for independent 'validation' of subgroup results. However in clinical trials investigating novel therapies/indications there may not be directly relevant data sources available. The consideration of observational (e.g., real-world data) sources could be helpful; see* Wang *et al*[30,31] *for a recent example and methods for utilizing insurance claims databases. While not an independent (external) evaluation, the proposed cross-validation assessments provide some (internal) diagnostic value.*

The subgroup findings from our analyses of the breast cancer and HIV trials could inform patient consultation. In the breast cancer trial, comparing hormonal therapy (tamoxifen) to chemotherapy, the estimated (bias-corrected) hazard ratio for subjects with positive estrogen levels (representing approximately 88% *of the study population) was* 0.64 *(*0.44, 0.93*) which suggests a slightly stronger benefit relative to the ITT population (*0.64 *vs* 0.69*). Though not dramatic, this could increase patients' confidence ("relative to ITT"); in contrast, patients with zero estrogen may want to consider alternatives (We note tamoxifen with low levels of estrogen seems controversial.*[21,22]*). In the HIV trial, comparing the combination of zidovudine and didanosine to monotherapy didanosine, the benefiting subgroup was generally comprised of subjects who were treatment naive or who had recent zidovudine use (within 30 days of study treatment initiation) but with less than a year of prior antiretroviral therapy. For these subjects, the estimated hazard ratio of* 0.59 *(*0.37, 0.94*) was relatively more substantial compared to the estimated hazard ratio of* 0.84 *for the ITT population. In terms of future trials, these findings could inform study designs such as inclusion criteria (e.g., consider excluding subjects with zero estrogen levels from tamoxifen trials) and/or randomization stratification factors, as well as testing strategies (e.g., pre-specify testing in the [zidovudine plus didanosine] benefiting subgroup described above). However, in general, we would caution against extrapolating findings to comparisons of regimens besides the control regimens that were studied in the trials.*

Subgroup analyses in Phase 2 trials can be the most actionable and impactful to inform Phase 3 study designs and analyses including: testing strategy; randomization stratification factors; and forest plot subgroup specifications. In particular, if there exists a sub-population that could potentially be harmed then identification in Phase 2 could mitigate the risk in later development (e.g., by implementing exclusion criteria/recommendations). Realistically, only substantial heterogeneous treatment effects can be identified and well estimated in Phase 2 settings; nevertheless, our simulation results under models M_2 *(*$N = 500$*) and* M_3 *(*$N = 300$*) suggests potential. On the other hand, in Phase 3 registrational trials the pre-specified subgroup (forest plot) results may suggest potential lack-of-benefit in a subgroup.*[32] *A comprehensive evaluation of subgroups, targeting large effects, may reveal a more accurate characterization than the pre-specified subgroups. Additionally, in multi-regional clinical trials the establishment of consistency*[33] *can be challenging; the identification of marked subgroup effects in the global trial could inform the evaluation of independent regional trials.*

2.5 Question 5

Related to the previous comment, perhaps this is beyond the current scope of the paper, but it might also be good to include some discussion of how to validate findings of this type of analysis, where the goal is identifying subgroups consistent with harm. It's one thing to prospectively validate a finding that a subgroup should do well (such as those targeted in the original virtual twins paper), but if the patients identified may actually be harmed by the experimental therapy, it seems like it might not be ethical to do prospective validation. Do the authors have a sense of how they might go about validating their findings in practice? Would something like sample splitting (say, before any steps of their procedure had been applied) work if the data set were large enough?

Thank you for your comments. As mentioned in responses above, we were motivated by your comment to implement two forms of cross-validation that we illustrated in our applications. We believe cross-validation is a useful 'internal diagnostic', but of course cannot generally serve as an external validation evaluation. If a sub-population is identified as being potentially harmed we would not envision prospectively testing this in a new trial for the reasons you mention. Rather, we would propose considering excluding such subjects from future trials if the level of evidence warranted. From a regulatory perspective, Amatya et al[32] discuss regulatory examples where some indications were restricted based on subgroup findings that were considered biologically plausible.

In the previous response we touch on these aspects (Sections form the revised discussion):

... In our data analysis applications we consider available data sources to evaluate the plausibility of our subgroup findings. Patient-level meta-analyses of randomized trials[24,28] *seems the most feasible and robust avenue for independent 'validation' of subgroup results. However in clinical trials investigating novel therapies/indications there may not be directly relevant data sources available. The consideration of observational (e.g., real-world data) sources could be helpful; see Wang et al*[30,31] *for a recent example and methods for utilizing insurance claims databases. While not an independent (external) evaluation, the proposed cross-validation assessments provide some (internal) diagnostic value.*

The subgroup findings from our analyses of the breast cancer and HIV trials could inform patient consultation. In the breast cancer trial, comparing hormonal therapy (tamoxifen) to chemotherapy, the estimated (bias-corrected) hazard ratio for subjects with positive estrogen levels (representing approximately 88% *of the study population) was* 0.64 (0.44, 0.93) *which suggests a slightly stronger benefit relative to the ITT population (*0.64 *vs* 0.69*). Though not dramatic, this could increase patients' confidence ("relative to ITT"); in contrast, patients with zero estrogen may want to consider alternatives (We note tamoxifen with low levels of estrogen seems controversial*[21,22]*.). In the HIV trial, comparing the combination of zidovudine and didanosine to monotherapy didanosine, the benefiting subgroup was generally comprised of subjects who were treatment naive or who had recent zidovudine use (within 30 days of study treatment initiation) but with less than a year of prior antiretroviral therapy. For these subjects, the estimated hazard ratio of* 0.59 (0.37, 0.94) *was relatively more substantial compared to the estimated hazard ratio of* 0.84 *for the ITT population. In terms of future trials, these findings could inform study designs such as inclusion criteria (e.g., consider excluding subjects with zero estrogen levels from tamoxifen trials) and/or randomization stratification factors, as well as testing strategies (e.g., pre-specify testing in the [zidovudine plus didanosine] benefiting subgroup described above). However, in general, we would caution against extrapolating findings to comparisons of regimens besides the control regimens that were studied in the trials.*

Subgroup analyses in Phase 2 trials can be the most actionable and impactful to inform Phase 3 study designs and analyses including: testing strategy; randomization stratification factors; and forest plot subgroup specifications. In particular, if there exists a sub-population that could potentially be harmed then identification in Phase 2 could mitigate the risk in later development (e.g., by implementing exclusion criteria/recommendations). Realistically, only substantial heterogeneous treatment effects can be identified and well estimated in Phase 2 settings; nevertheless, our simulation results under models M_2 *(*$N = 500$*) and* M_3 *(*$N = 300$*) suggests potential. On the other hand, in Phase 3 registrational trials the pre-specified subgroup (forest plot) results may suggest potential lack-of-benefit in a subgroup.*[32] *A comprehensive evaluation of subgroups, targeting large effects, may reveal a more accurate characterization than the pre-specified subgroups. Additionally, in multi-regional clinical trials the establishment of consistency*[33] *can be challenging; the identification of marked subgroup effects in the global trial could inform the evaluation of independent regional trials.*

References

1. Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics* 2019; 47(2): 1148–1178.

2. Athey S, Wager S. Policy learning with observational data. *Econometrica* 2021; 89(1): 133–161.

3. 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.

4. Simon N, Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* 2011; 39(5): 113.

5. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011; 30(24): 2867-2880.

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

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

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

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

10. Sverdrup E, Kanodia A, Zhou Z, Athey S, Wager S. policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization over Trees. 2023. R package version 1.2.2.

11. Aalen O, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect?. *Lifetime data analysis* 2015: 579593.

12. Agboola F, Whittington MD, Pearson SD. Advancing Health Technology Assessment Methods that Support Health Equity. Institute for Clinical and Economic Review. March 15, 2023. https://icer.org/assessment/health-technology-assessment-methods-that-support-health-equity-2023.

13. 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.

14. Loh WY, Man M, Wang S. Subgroups from regression trees with adjustment for prognostic effects and postselection inference. *Statistics in medicine* 2018; 38(4): 545–557.

15. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 2004; 75(1): 45-49.

16. Guo X, He X. Inference on Selected Subgroups in Clinical Trials. *Journal of the American Statistical Association* 2021; 116(535): 1498-1506.

17. Knaus MC. Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal* 2022; 25(3): 602-627.

18. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying Important Risk Factors for Survival in Patient With Systolic Heart Failure Using Random Survival Forests. *Circulation: Cardiovascular Quality and Outcomes* 2011; 4(1): 39-45.

19. Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. *Ann. Appl. Statist.* 2008; 2(3): 841–860.

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

21. Manna S, Holz MK. Tamoxifen Action in ER-Negative Breast Cancer. *Signal Transduction Insights* 2016; 5: STI.S29901.

22. Yu KD, Cai YW, Wu SY, Shui RH, Shao ZM. Estrogen receptor-low breast cancer: Biology chaos and treatment paradox. *Cancer Communications* 2021; 41(10): 968-980.

23. Cuzick J, Powles T, Veronesi U, et al. Overview of the main outcomes in breast-cancer prevention trials. *The Lancet* 2003; 361(9354): 296-300.

24. Early Breast Cancer Trialists' Collaborative Group (EBCTCG) . Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The Lancet* 2011; 378(9793): 771-784.

25. Foekens JA, Peters HA, Look MP, et al. The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients1. *Cancer Research* 2000; 60(3): 636-643.

26. Royston P, Altman D. External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology* 2013.

27. Katzenstein D, Hughes M, Albrecht M, et al. Virologic and CD4 Cell Response to Zidovudine or Zidovudine and Lamivudine Following Didanosine Treatment of Human Immunodeficiency Virus Infection. *AIDS Research and Human Retroviruses* 2001; 17(3): 203-210.

28. HIV Trialists' Collaborative Group . Zidovudine, didanosine, and zalcitabine in the treatment of HIV infection: meta-analyses of the randomised e vidence. *The Lancet* 1999; 353(9169): 2014-2025.

29. 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.

30. Wang SV, Schneeweiss S, Initiative RD. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA* 2023; 329(16): 1376-1385.

31. Sheldrick RC. Randomized Trials vs Real-world Evidence: How Can Both Inform Decision-making?. *JAMA* 2023; 329(16): 1352-1353.

32. 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.

33. Ying L, Song F, Chow SC, et al. On evaluation of consistency in multi-regional clinical trials. *Journal of Biopharmaceutical Statistics* 2018; 28(5): 840–856.

□