The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2266–2278
https://doi.org/10.1214/21-AOAS1589

AN OMNIBUS TEST FOR DETECTION OF SUBGROUP TREATMENT EFFECTS VIA DATA PARTITIONING

BY YIFEI SUN[1,a], XUMING HE[2,c] AND JIANHUA HU[1,b]

[1]*Department of Biostatistics, Columbia University,* [a]*ys3072@cumc.columbia.edu,* [b]*jh3992@cumc.columbia.edu*
[2]*Department of Statistics, University of Michigan,* [c]*xmhe@umich.edu*

Late-stage clinical trials have been conducted primarily to establish the efficacy of a new treatment in an intended population. A corollary of population heterogeneity in clinical trials is that a treatment might be effective for one or more subgroups, rather than for the whole population of interest. As an example, the phase III clinical trial of panitumumab in metastatic colorectal cancer patients failed to demonstrate its efficacy in the overall population, but a subgroup associated with tumor KRAS status was found to be promising (Peeters et al. (*Am. J. Clin. Oncol.* **28** (2010) 4706–4713)). As we search for such subgroups via data partitioning based on a large number of biomarkers, we need to guard against inflated type I error rates due to multiple testing. Commonly-used multiplicity adjustments tend to lose power for the detection of subgroup treatment effects. We develop an effective omnibus test to detect the existence of, at least, one subgroup treatment effect, allowing a large number of possible subgroups to be considered and possibly censored outcomes. Applied to the panitumumab trial data, the proposed test would confirm a significant subgroup treatment effect. Empirical studies also show that the proposed test is applicable to a variety of outcome variables and maintains robust statistical power.

1. Introduction. The primary purpose of a late-stage clinical trial is to establish efficacy of a new treatment measured by the overall treatment effect in an intended population. However, it is quite common that treatment effects are sufficiently heterogeneous across subgroups such that a treatment might be most effective for one subgroup of patients when the overall treatment effect is marginal. Identification of subgroups in clinical settings is routinely carried out with seemingly promising findings. For example, eight-week treatment with ledipasvir/sofosbuvir was found to be effective for chronic hepatitis C subjects with certain genotypes (O'Brien, Kuhs and Pfeiffer (2014)). The medicine Vectibix (panitumumab) for metastatic colorectal cancer was approved for patients without RAS gene mutation when the overall treatment effect was much less promising (Peeters et al. (2010, 2015)).

In this paper we focus on the randomized phase III study of panitumumab in metastatic colorectal cancer (mCRC) patients (Peeters et al. (2010)). Panitumumab is a fully human antiepidermal growth factor receptor (EGFR) monoclonal antibody that improves progression-free survival in chemotherapy-refractory mCRC. This trial evaluated the efficacy and safety of panitumumab plus fluorouracil, leucovorin, and irinotecan (FOLFIRI), compared with FOLFIRI alone, after failure of the initial treatment for mCRC. Besides the treatment assignment, the data contain other patient characteristics such as age, sex, and Kirsten rat sarcoma virus (KRAS) gene mutation status. The primary endpoint is the progression free survival, and the secondary endpoint is the binary clinical response. Both endpoints were considered in our analysis.

In the aforementioned panitumumab trial, given that the treatment did not pass the regulatory approval for the general population, investigators naturally asked if there is any patient

Received May 2021; revised October 2021.
Key words and phrases. Bootstrap, data partitioning, clinical trials, high-dimensional covariates, subgroup treatment effect.

subgroup that benefits from the treatment more significantly. To this end, attempts have been made to explore the potential for further refinement of patient selection using data from the original study (Peeters et al. (2015)). It is well recognized that, if the same trial data are used to identify subgroups and to evaluate treatment effects for the selected subgroup, we must account for the selection bias in the post hoc analysis; see Naggara et al. (2011) and Dmitrienko et al. (2020). When a large number of possible subgroups are searched over, simple multiplicity adjustments (e.g., the Bonferroni adjustment) often lead to a substantial loss of statistical power. An omnibus test on the treatment effects of subgroups would strengthen statistical support to further analysis and validation of certain subgroup(s) for the experimental treatment. Our goal is to develop a formal statistical test on the existence of favorable subgroups.

Several methods have been studied in the literature for identifying subgroups with differential treatment effects based on decision trees. For example, a generalized unbiased interaction detection and estimation algorithm was discussed in Loh (2009) and Loh, He and Man (2015). Coupled with the extensive tree-split search, this method uses multiple significance tests on main effects and interactions to reduce the size of the tree structures and to improve precision of the splits. Sies, Demyttenaere and Mechelen (2019) focused on identifying subgroups with heterogeneous treatment effects and discussed performances of several existing methods. Ondra et al. (2016) reviewed various methods on identification of subgroups with heterogeneous treatment effects. Dmitrienko, Millen and Lipkovich (2017) reviewed multiplicity adjustment methods in confirmatory subgroup analysis based on prespecified patient subpopulations.

The evaluation of the subgroup effects has also attracted attention in the recent literature. Guo and He (2021) developed a debiasing resampling approach to the evaluation of the best selected subgroup without being (unnecessarily) protective against other subgroups. Wager and Athey (2018) developed a nonparametric causal forest for estimating heterogeneous treatment effects based on random forest algorithms. Joshi et al. (2019) discussed the problem of estimating a treatment subgroup, based on a single biomarker, and that of testing for treatment effect in the identified subgroup.

Statistical tests for the existence of heterogeneous or differential subgroups have limited coverage in the literature. Shen and He (2015) introduced a structured logistic-normal mixture model framework to perform a confirmatory statistical test for the existence of subgroups. Shen and Qu (2020) expanded the work to handle longitudinal data where the heterogeneous treatment effect is modeled as a random effect from a two-component mixture model. Behr et al. (2020) presented a statistical test on association between the response variable and all levels of the tree hierarchy based on given tree structures. These methods, however, focus on differential subgroup effects which is broader than the existence of any subgroup with significant treatment effects.

In the present paper we propose an omnibus test on the existence of a subgroup with favorable treatment effects. More specifically, we focus on subgroups that are induced through partitions based on covariates. The proposed test, supported by the recent theory on high-dimensional bootstrap-based tests of Chernozhukov, Chetverikov and Kato (2013) and Chernozhukov, Chetverikov and Kato (2019), permits a large number of possible partitions and works with common measures of treatment effects with various types of outcome variables. The proposed method allows a flexible form of covariate adjustments in the calculation of treatment effects and works with high dimensional covariates as well. Simulation studies demonstrate that the proposed method preserves type-I error rates well in a variety of scenarios.

The rest of the paper is organized as follows. Section 2 describes the proposed framework and the construction of a nonparametric omnibus test for subgroup treatment effects using

the bootstrap. Section 3 provides an extension of the proposed method to censored survival outcomes. Section 4 presents comparative simulation studies, and Section 5 revisits the phase III trial data on panitumumab to see what we can learn from the proposed test. We provide some concluding remarks in Section 6.

2. Method. Let Y be the outcome of interest, and assume without loss of generality that a greater value of Y is a more favorable outcome. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a p-dimensional vector of covariates. The covariates may include demographic factors, biomarkers, and their low-order interactions. Denote by T the binary treatment indicator, where $T = 1$ indicates treatment and $T = 0$ indicates control. In this paper we focus on randomized trials where the treatment assignment T is independent of the covariates $\mathbf{X}$. We consider testing the null hypothesis of no treatment effect in any subpopulation, based on the observed data $\{(\mathbf{X}_i, Y_i, T_i),\ i = 1, \ldots, n\}$, as independent realizations from the joint probability distribution of $(\mathbf{X}, Y, T)$.

2.1. *The proposed test statistic.* We consider J potential subgroups of interest, $\mathcal{G}_j$, where $j = 1, \ldots, J$. Each of them is a subset of the sample space of $\mathbf{X}$ and may take various forms. For example, $\{\mathbf{X} : X_k > c_k\}$ for some cut-off points c_k $(k = 1, \ldots, p)$ and their complements can form subgroups. The entire population is also considered as a subgroup of interest. The total number of subgroups J could be greater than p. Our null hypothesis of interest is

H_0: the conditional distribution of Y, given $(\mathbf{X} \in \mathcal{G}_j, T)$, does not vary with T for any j.

To reduce the variability in the outcome variable, we consider adjusting for $\mathbf{X}$ by subtracting a function of $\mathbf{X}$, denoted by f, from Y and use $Y - f(\mathbf{X})$ as the adjusted outcome. A specific form of f will be described later. We then rely on an estimated treatment effect on the adjusted outcome for subgroup j,

$$\begin{aligned}\widehat{\beta}_j = {} & \widehat{E}\big[\{Y - f(\mathbf{X})\} I(T = 1, \mathbf{X} \in \mathcal{G}_j)\big] \widehat{E}\big\{I(T = 0, \mathbf{X} \in \mathcal{G}_j)\big\} \\ & - \widehat{E}\big[\{Y - f(\mathbf{X})\} I(T = 0, \mathbf{X} \in \mathcal{G}_j)\big] \widehat{E}\big\{I(T = 1, \mathbf{X} \in \mathcal{G}_j)\big\},\end{aligned} \tag{1}$$

where $\widehat{E}$ denotes the empirical expectation based on the observed data. Note that $\widehat{\beta}_j$ is an estimate of the scaled treatment effect

$$\beta_{j0} = c_j\big(E\big[\{Y - f(\mathbf{X})\} \mid T = 1, \mathbf{X} \in \mathcal{G}_j\big] - E\big[\{Y - f(\mathbf{X})\} \mid T = 0, \mathbf{X} \in \mathcal{G}_j\big]\big),$$

where $c_j = P(T = 0, \mathbf{X} \in \mathcal{G}_j) P(T = 1, \mathbf{X} \in \mathcal{G}_j)$. The reason we use the scaled quantity here is to avoid instability issues in the estimation of the conditional expectations when few cases are under the treatment in one of the subgroups. Since T is independent of $\mathbf{X}$ in randomized trials, we also have $\beta_{j0} = c_j\{E(Y \mid T = 1, \mathbf{X} \in \mathcal{G}_j) - E(Y \mid T = 0, \mathbf{X} \in \mathcal{G}_j)\}$ which is free of f. Under the null hypothesis we have $\beta_{j0} = 0$.

Note that β_{j0} can be interpreted as a scaled difference in the conditional expectations for subgroup j, and more importantly, $\widehat{\beta}_j$ is a convenient summary statistic for comparison of the conditional distributions. Since the mean can be expressed as an integral of the survival function, β_{j0} can be viewed as the integral of the difference in two survival functions scaled by c_j. One may also consider a more general class of statistics for testing equality of distributions, defined as the integral of the weighted difference in the estimated survival functions of the treated and control subjects in group j (see, e.g., Pepe and Fleming (1991)). Another possible type of statistics is discussed in Section 3. For simplicity, we focus on $\widehat{\beta}_j$ defined in (1), and similar arguments apply to many other test statistics; each of which would aim to detect a specific form of the deviations from the null hypothesis.

Suppose n_j/n converges to a constant, and β_{j0} is the limiting value of $\widehat{\beta}_j$ as $n \to \infty$. The statistic $\sqrt{n}(\widehat{\beta}_j - \beta_{j0})$ has an approximately linear form, $\sqrt{n}(\widehat{\beta}_j - \beta_{j0}) = n^{-1/2} \sum_{i=1}^{n} \phi_{ji} +$

r_{jn}, where ϕ_{ji} are zero mean random variables (with finite variance) and r_{jn} is a higher-order term that converges to zero. Hence, $\sqrt{n}(\widehat{\beta_j} - \beta_{j0})$ converges in distribution to $N(0, \sigma_j^2)$, where $\sigma_j^2 = E\phi_{ji}^2$. If the covariates have heavy tails or if the estimator is an extreme statistic or a mode estimator in nature, its limiting distribution may become irregular and fall out of our considerations here. With $\widehat{\beta}_j$ defined in (1), the influence function ϕ_{ji} has the form $\phi_{ji} = \{Y_i - f(\mathbf{X}_i)\}I(T_i = 1, \mathbf{X}_i \in \mathcal{G}_j)E\{I(T = 0, \mathbf{X} \in \mathcal{G}_j)\} + E[\{Y - f(\mathbf{X})\}I(T = 1, \mathbf{X} \in \mathcal{G}_j)]I(T_i = 0, \mathbf{X}_i \in \mathcal{G}_j) - \{Y_i - f(\mathbf{X}_i)\}I(T_i = 0, \mathbf{X}_i \in \mathcal{G}_j)E\{I(T = 1, \mathbf{X} \in \mathcal{G}_j)\} - E[\{Y - f(\mathbf{X})\}I(T = 0, \mathbf{X} \in \mathcal{G}_j)]I(T_i = 1, \mathbf{X}_i \in \mathcal{G}_j) - 2\beta_{j0}$. Suppose $\widehat{\sigma}_j$ is an estimator for σ_j under the null hypothesis of no treatment effect in any subgroup. In practice, $\widehat{\sigma}_j$ can be obtained using the bootstrap method, as discussed in Section 2.3. We propose to use the statistic

$$T_n = \sqrt{n} \sup_j (\widehat{\beta}_j / \widehat{\sigma}_j).$$

Under the null hypothesis the distribution of T_n is approximated by the distribution of the supremum of J zero-mean normally distributed random variables, as shown by Chernozhukov, Chetverikov and Kato (2013) for high-dimensional supremum-type statistics. We reject the null hypothesis at a given level of significance if T_n exceeds the critical value to be determined later.

2.2. *Implementation of the test.* The proposed test described above works for various choices of f. A simple choice would be $f(\mathbf{X}) = 0$ for all $\mathbf{X}$ so that the outcome is not adjusted. We find that the test tends to have better finite-sample power when σ_j^2 are smaller under H_0, and the suggested use of $f(\mathbf{X})$ is to make the resulting σ_j^2's more favorable for the test.

To explain how the adjustment based on f improves power, we take, for example, a data generative model $Y = f_0(\mathbf{X}) + \gamma_0 T + \epsilon$, where ϵ represents an independent noise. If we average the outcome Y over all the subjects in subgroup j for $T = 1$ or $T = 0$, the inherent variability in $\widehat{\beta}_j$ consists of those of $\mathbf{X}$ in subgroup j and ϵ. The variances σ_j^2 would be reduced if $f(\mathbf{X})$ captures part of $f_0(\mathbf{X})$. Of course, it would be ideal to set $f = f_0$ in this stylistic case, but this would be unrealistic in real applications.

In practice, the function f can be estimated from the data. To strike a balance between simplicity and efficacy, we regress Y on a prespecified subset of variables in $\mathbf{X}$, denoted by $s(\mathbf{X})$. More specifically, we take $\widehat{f}(\mathbf{X}) = \widehat{\boldsymbol{\theta}}^\top s(\mathbf{X})$, where $\widehat{\boldsymbol{\theta}}$ is obtained from minimizing $\sum_{i=1}^n (Y_i - \theta_0 - \boldsymbol{\theta}^\top s(\mathbf{X}_i))^2$. When p is small, we use $s(\mathbf{X}) = \mathbf{X}$. When p is large, one can choose a small subset of predictive covariates based on domain knowledge. A data-driven approach that choose $s(\mathbf{X})$, based on adaptive lasso (Zou (2006)), was investigated in our simulation studies and yielded reasonably good performances. The relevant asymptotic behavior of the adaptive lasso estimator can be found in Lu, Goldberg and Fine (2012). Our proposed method does not require "correctly" selecting all the predictive variables to guarantee the type I error rate, but sensible choices of $s(\mathbf{X})$ may yield better powers of the test. Our operational estimate for subgroup j is given as

$$\begin{aligned}\widehat{\beta}_j = {} & \widehat{E}\big[\{Y - \widehat{f}(\mathbf{X})\}I(T = 1, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}\big\{I(T = 0, \mathbf{X} \in \mathcal{G}_j)\big\} \\ & - \widehat{E}\big[\{Y - \widehat{f}(\mathbf{X})\}I(T = 0, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}\big\{I(T = 1, \mathbf{X} \in \mathcal{G}_j)\big\}.\end{aligned} \tag{2}$$

We note that the underlying model of Y, given $\mathbf{X}$, in the absence of treatment is unknown and not necessarily linear. However, the least squares objective function, $E(Y - \theta_0 - \boldsymbol{\theta}^\top s(\mathbf{X}))^2$, has a unique minimizer $(\theta_0^*, \boldsymbol{\theta}^*)$ as long as the distributions of $\mathbf{X}$ and Y are not degenerate. Then, $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges to a zero mean normal random vector, and the estimated function $\widehat{f}(\boldsymbol{x})$ converges to $f(\boldsymbol{x}) = \boldsymbol{\theta}^{*\top} s(\boldsymbol{x})$ in the limit. Under the null hypothesis of

no treatment effect and the assumption that T is independent of $\mathbf{X}$, the influence function ϕ_{ji} for $\widehat{\beta}_j$ in (2) is the same as that when f is known, because the variability, due to estimating f, in the first and second terms in $\widehat{\beta}_j$ cancel out.

The distribution of T_n under the null hypothesis does not have a Gaussian limit, but as a supremum statistic over J subgroups, it has been investigated in Chernozhukov, Chetverikov and Kato (2013), Chernozhukov, Chetverikov and Kato (2017). As long as $\widehat{\beta}_j$ is asymptotically linear, we know from the literature that T_n is bootstrap-able, even as J increases subexponentially with n. For this reason we propose to use the bootstrap to calculate the critical value of T_n.

2.3. *Use of the bootstrap*. A bootstrap sample is a random sample of n subjects drawn with replacement from the original sample. We shall use $\widehat{E}^*$ to denote the empirical mean and $\widehat{f}^*$ to denote the estimate of f based on the bootstrap sample. Define

$$\begin{aligned}\widehat{\beta}_j^* = {} & \widehat{E}^*\big[\{Y - \widehat{f}^*(\mathbf{X})\}I(T=1, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}^*\big\{I(T=0, \mathbf{X} \in \mathcal{G}_j)\big\} \\ & - \widehat{E}^*\big[\{Y - \widehat{f}^*(\mathbf{X})\}I(T=0, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}^*\big\{I(T=1, \mathbf{X} \in \mathcal{G}_j)\big\}\end{aligned}$$

which replace the empirical estimates in (2) with the estimates from the bootstrap sample. The standard deviation of $\sqrt{n}\widehat{\beta}_j^*$, computed from a sufficiently large number of $\widehat{\beta}_j^*$, consistently estimates σ_j and may be a possible choice for $\widehat{\sigma}_j$. Another choice of $\widehat{\sigma}_j$ is given below.

Although the bootstrap distribution of $\sqrt{n}\sup_j(\widehat{\beta}_j^* - \widehat{\beta}_j)/\widehat{\sigma}_j$ approximates the sampling distribution of $\sqrt{n}\sup_j(\widehat{\beta}_j/\widehat{\sigma}_j)$ under the null hypothesis and thus leads to approximate validity of the proposed test, we cannot say much about the power when the null hypothesis is not true for the original data. We note that the power of the test can be potentially improved if we replace Y by a globally adjusted outcome $\widetilde{Y} = Y - \widehat{g}(T, \mathbf{X})$, where $\widehat{g}(T, \mathbf{X})$ is a regression estimate that approximates the treatment effects such that $\widehat{g}(T, \mathbf{X}) \approx 0$ under the null hypothesis. A convenient choice of the global adjustment takes the form

$$\widehat{g}(T, \mathbf{X}) = \widehat{\gamma}_0 T + \widehat{\boldsymbol{\gamma}}^\top s_2(\mathbf{X})T,$$

where $s_2(\mathbf{X})$ is a prespecified subset of covariates in $\mathbf{X}$, $\widehat{\gamma}_0$ and $\widehat{\boldsymbol{\gamma}}$ denote the estimated regression coefficients based on a working model $Y = \eta_0 + \boldsymbol{\eta}^\top s_2(\mathbf{X}) + \gamma_0 T + \boldsymbol{\gamma}^\top s_2(\mathbf{X})T + \epsilon$. In practice, $s_2(\mathbf{X})$ can include variables in $s(\mathbf{X})$ as well as other variables that potentially affect the treatment effect. Denoted by $(\eta_0^*, \boldsymbol{\eta}^*, \gamma_0^*, \boldsymbol{\gamma}^*)$ the minimizer of $E(Y - \eta_0 - \boldsymbol{\eta}^\top s_2(\mathbf{X}) - \gamma_0 T - \boldsymbol{\gamma}^\top s_2(\mathbf{X})T)^2$, then $\widehat{g}(t, \boldsymbol{x})$ converges to the function $g(t, \boldsymbol{x}) = \gamma_0^* t + \boldsymbol{\gamma}^{*\top} s_2(\boldsymbol{x})t$. Under the null hypothesis of no treatment effect, we have $\gamma_0^* = 0$ and $\boldsymbol{\gamma}^* = \mathbf{0}$.

We now summarize the proposed bootstrap implementation to determine the critical values of T_n. By working with $\{(\widetilde{Y}_i, \mathbf{X}_i, T_i), i = 1, \ldots, n\}$, we estimate the function f, as described in the preceding subsection with the resulting estimate $\widetilde{f}$. Then, we obtain the quantities

$$\begin{aligned}\widetilde{\beta}_j = {} & \widehat{E}\big[\{\widetilde{Y} - \widetilde{f}(\mathbf{X})\}I(T=1, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}\big\{I(T=0, \mathbf{X} \in \mathcal{G}_j)\big\} \\ & - \widehat{E}\big[\{\widetilde{Y} - \widetilde{f}(\mathbf{X})\}I(T=0, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}\big\{I(T=1, \mathbf{X} \in \mathcal{G}_j)\big\}, \\ \widetilde{\beta}_j^* = {} & \widehat{E}^*\big[\{\widetilde{Y} - \widetilde{f}^*(\mathbf{X})\}I(T=1, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}^*\big\{I(T=0, \mathbf{X} \in \mathcal{G}_j)\big\} \\ & - \widehat{E}^*\big[\{\widetilde{Y} - \widetilde{f}^*(\mathbf{X})\}I(T=0, \mathbf{X} \in \mathcal{G}_j)\big]\widehat{E}^*\big\{I(T=1, \mathbf{X} \in \mathcal{G}_j)\big\},\end{aligned}$$

where "*" refers to estimators constructed using the bootstrap samples. Let $\widetilde{\sigma}_j$ be the standard deviation of $\widetilde{\beta}_j^*$ calculated using a sufficiently large number of bootstrap samples. The bootstrap test statistic is given as $T_n^* = \sqrt{n}\sup_j(\widetilde{\beta}_j^* - \widetilde{\beta}_j)/\widetilde{\sigma}_j$. Under the null hypothesis, $\widetilde{\sigma}_j$ consistently estimates σ_j. Then, we can use $\widetilde{\sigma}_j$ in T_n and take $T_n = \sqrt{n}\sup_j(\widehat{\beta}_j/\widetilde{\sigma}_j)$. On the other hand, when the null hypothesis is not true, $\widetilde{\sigma}_j$ can be quite different from the standard

deviation of $\widehat{\beta}_j^*$. Under the null hypothesis the bootstrap sampling distribution of T_n^* approximates the sampling distribution of T_n. The $1-\alpha$ quantile of T_n^* is used as the critical value for T_n at the significance level α.

REMARK. Note that the bootstrap procedure uses $\widehat{g}(T, \mathbf{X})$ to reduce the possible treatment effects, and the proposed test uses $\widetilde{f}(\mathbf{X})$ and $\widetilde{f}^*(\mathbf{X})$ to adjust the outcome in the original sample and the bootstrap sample, respectively. If the true model takes the form $Y = f_0(\mathbf{X}) + f_1(\mathbf{X})T + \epsilon$, where ϵ is an independent error, the proposed adjustment $\widehat{g}$ can be viewed as removing some variability, due to $f_1(\mathbf{X})T$, so that the bootstrap distribution after the adjustment can better approximate the distribution of T_n were there no treatment effects. However, the validity of the proposed test (under the null hypothesis) does not require the specific forms of f and g to reflect the true data generation model, and only the finite-sample power of the test varies with the adjustments.

3. Extension to survival data. Consider the outcome Y as a survival time of interest which is common in clinical studies. In practice, the survival time is often subject to right censoring due to loss to follow-up or natural study termination. Let C be the censoring time, $Z = \min(Y, C)$ the follow-up time, and $\Delta = I(Y \le C)$ the event indicator. We assume C is independent of $(Y, \mathbf{X}, T)$, where $\mathbf{X} \in \mathbb{R}^p$ is the covariate and T is the treatment indicator, as in the previous section. We also assume T is independent of $\mathbf{X}$. We consider testing the null hypothesis of no treatment effect on Y in any subpopulation, based on the observed data $\{(\mathbf{X}_i, Z_i, \Delta_i, T_i),\ i = 1, \ldots, n\}$, which are independent realizations from the joint probability distribution of $(\mathbf{X}, Z, \Delta, T)$.

For right-censored data, a natural choice for $\widehat{\beta}_j$ is the weighted log-rank test statistic. We also suggest using the adjusted outcome $Y \exp\{-f(\mathbf{X})\}$, instead of Y, in constructing the test statistic. The function f can be estimated from a working accelerated failure time (AFT) model $\log Y = \boldsymbol{\theta}^\top s(\mathbf{X}) + \epsilon$, where $s(\mathbf{X})$ is a prespecified subset of the variables in $\mathbf{X}$ and ϵ is a random error. We set $\widehat{f}(\mathbf{X}) = \widehat{\boldsymbol{\theta}}^\top s(\mathbf{X})$, where $\widehat{\boldsymbol{\theta}}$ are obtained by solving the weighted log-rank estimating equations (Tsiatis (1990)). Denote by $\boldsymbol{\theta}^*$ the solution of the limiting estimating equation, then $\widehat{f}(\boldsymbol{x})$ converges to the function $f(\boldsymbol{x}) = \boldsymbol{\theta}^{*\top} s(\boldsymbol{x})$. One may also consider other choices of f as appropriate. It is worth pointing out, as discussed in Section 2, the validity of our test does not require the working model to be correctly specified.

Define the counting process $N(u; \widehat{f}) = \Delta I(Z \exp\{-\widehat{f}(\mathbf{X})\} \le u)$ and the at-risk process $R(u; \widehat{f}) = I(Z \exp\{-\widehat{f}(\mathbf{X})\} \ge u)$. The treatment effect in subgroup j can be quantified by

$$\begin{aligned}\widehat{\beta}_j = \int_0^\infty \big[&\widehat{E}\{I(T=1)R(u;\widehat{f}) \mid \mathbf{X} \in \mathcal{G}_j\}\widehat{E}\{dN(u;\widehat{f}) \mid \mathbf{X} \in \mathcal{G}_j\} \\ &- \widehat{E}\{R(u;\widehat{f}) \mid \mathbf{X} \in \mathcal{G}_j\}\widehat{E}\{I(T=1)\,dN(u;\widehat{f}) \mid \mathbf{X} \in \mathcal{G}_j\}\big].\end{aligned} \tag{3}$$

The statistic $\widehat{\beta}_j$ can be viewed as an analogue of the weighted log-rank test statistic with the Gehan weight (Gehan (1965)). As $n \to \infty$, $\widehat{\beta}_j$ converges to its limiting value β_{j0} with

$$\begin{aligned}\beta_{j0} = {} & P(\mathcal{Y}_1 > \mathcal{Y}_2, T_1 = 1, T_2 = 0, \min(\mathcal{Y}_1, \mathcal{Y}_2) \le \min(\mathcal{C}_1, \mathcal{C}_2) \mid \mathbf{X}_1 \in \mathcal{G}_j, \mathbf{X}_2 \in \mathcal{G}_j) \\ & - P(\mathcal{Y}_2 > \mathcal{Y}_1, T_1 = 1, T_2 = 0, \min(\mathcal{Y}_1, \mathcal{Y}_2) \le \min(\mathcal{C}_1, \mathcal{C}_2) \mid \mathbf{X}_1 \in \mathcal{G}_j, \mathbf{X}_2 \in \mathcal{G}_j),\end{aligned}$$

where we denote by $\mathcal{Y}_i = Y_i \exp\{-f(\mathbf{X}_i)\}$ and $\mathcal{C}_i = C_i \exp\{-f(\mathbf{X}_i)\}$ the adjusted survival time and censoring time, respectively, and $\{(\mathbf{X}_i, T_i, Y_i, C_i), i = 1, 2\}$ are independent realizations of $(\mathbf{X}, T, Y, C)$. In the absence of treatment effect, we have $\beta_{j0} = 0$. When the treatment prolongs the survival in the jth subgroup, β_{j0} is generally positive. We define the test statistic as $T_n = \sqrt{n} \sup_j (\widehat{\beta}_j / \widetilde{\sigma}_j)$, where $\widetilde{\sigma}_j$ is the standard error estimate of $\widehat{\beta}_j$ under the null hypothesis and is calculated using bootstrap as discussed below.

Similar to Section 2, we may also consider using the outcome $Y\exp\{-g(T,\mathbf{X})\}$ in the bootstrap procedure. For example, $\widehat{g}(T,\mathbf{X})=\widehat{\gamma}_0 T+\widehat{\boldsymbol{\gamma}}^\top s_2(\mathbf{X})T$ can be estimated from another working AFT model $\log Y=\boldsymbol{\theta}^\top s_2(\mathbf{X})+\gamma_0 T+\boldsymbol{\gamma}^\top s_2(\mathbf{X})T+\epsilon$, where $s_2(\mathbf{X})$ is a prespecified subset of covariates. When p is large, covariates in $s_2(\mathbf{X})$ can be selected based on marginal screening. Define $\widetilde{Z}=Z\exp\{-\widehat{g}(T,\mathbf{X})\}$, $\widetilde{N}(u;f)=\Delta I(\widetilde{Z}\exp\{-f(\mathbf{X})\}\le u)$, and $\widetilde{R}(u;f)=I(\widetilde{Z}\exp\{-f(\mathbf{X})\}\ge u)$. Using $\{(\mathbf{X}_i,\widetilde{Z}_i,\Delta_i,T_i),i=1,\ldots,n\}$, we estimate the function f, as described above, with Z replaced by $\widetilde{Z}$ and denote the resulting estimate by $\widetilde{f}$. On a bootstrap sample $\{(\mathbf{X}_i^*,\widetilde{Z}_i^*,\Delta_i^*,T_i^*),i=1,\ldots,n\}$, we estimate f in the same way and denote the estimator by $\widetilde{f}^*$. Then, we obtain the quantities

$$\begin{aligned}\widetilde{\beta}_j=\int_0^\infty&\big[\widehat{E}\{I(T=1)\widetilde{R}(u;\widetilde{f})\mid\mathbf{X}\in\mathcal{G}_j\}\widehat{E}\{d\widetilde{N}(u;\widetilde{f})\mid\mathbf{X}\in\mathcal{G}_j\}\\&-\widehat{E}\{\widetilde{R}(u;\widetilde{f})\mid\mathbf{X}\in\mathcal{G}_j\}\widehat{E}\{I(T=1)d\widetilde{N}(u;\widetilde{f})\mid\mathbf{X}\in\mathcal{G}_j\}\big],\\\widetilde{\beta}_j^*=\int_0^\infty&\big[\widehat{E}^*\{I(T=1)\widetilde{R}(u;\widetilde{f}^*)\mid\mathbf{X}\in\mathcal{G}_j\}\widehat{E}^*\{d\widetilde{N}(u;\widetilde{f}^*)\mid\mathbf{X}\in\mathcal{G}_j\}\\&-\widehat{E}^*\{\widetilde{R}(u;\widetilde{f}^*)\mid\mathbf{X}\in\mathcal{G}_j\}\widehat{E}^*\{I(T=1)d\widetilde{N}(u;\widetilde{f}^*)\mid\mathbf{X}\in\mathcal{G}_j\}\big].\end{aligned}$$

The bootstrap test statistic is given as $T_n^*=\sqrt{n}\sup_j(\widetilde{\beta}_j^*-\widetilde{\beta}_j)/\widetilde{\sigma}_j$, where $\widetilde{\sigma}_j$ is the standard deviation of $\sqrt{n}\widetilde{\beta}_j^*$ calculated using a sufficiently large number of bootstrap samples. The $1-\alpha$ quantile of T_n^* is used as the critical value for T_n at the significance level α.

4. Simulation. In this section we report simulation results to evaluate the performance of the proposed test relative to some of the obvious alternatives. In the first set of simulations, we considered uncensored data. The vector $\mathbf{X}=(X_1,\ldots,X_{50})$ contains 50 independent covariates. For $k=1,\ldots,10$ and $k=41,\ldots,50$, X_k was generated from a uniform distribution on $[-1,1]$; for $k=11,\ldots,20$ and $k=31,\ldots,40$, X_k was generated from a discrete probability distribution with $P(X_k=-1)=P(X_k=1)=0.5$; for $k=21,\ldots,30$, X_k was generated from the normal distribution $N(0,0.5^2)$. The treatment indicator T was generated from a Bernoulli distribution with success probability 0.5. The outcome Y were generated from the following models:

(I) $Y=X_1-X_{50}+aT\cdot\xi(\mathbf{X})+\epsilon$, $\xi(\mathbf{X})=I(X_1>0$ and $X_{25}>0)$;
(II) $Y=X_1-X_{50}+aT\cdot\xi(\mathbf{X})+\epsilon$, $\xi(\mathbf{X})=(X_1+X_{24}-X_{25}-X_{50}+0.5)/2$;
(III) $Y=I(X_1>0)-I(X_{50}>0)+aT\cdot\xi(\mathbf{X})+\epsilon$, $\xi(\mathbf{X})=(X_1^2+X_{24}+X_{25}+X_{50})/2$;
(IV) $Y=X_1X_{50}+aT\cdot\xi(\mathbf{X})+\epsilon$, $\xi(\mathbf{X})=(X_1+X_{25}^2/2)/2$;
(V) $Y=\exp\{X_1-X_{25}+aT\cdot\xi(\mathbf{X})\}+\epsilon$, $\xi(\mathbf{X})=I(X_{50}>0)-0.5$;
(VI) $Y=\exp\{X_1-X_{25}+aT\cdot\xi(\mathbf{X})\}+\epsilon$, $\xi(\mathbf{X})=I(X_1>0$ and $X_{25}>0)$.

The random error ϵ was generated from a normal distribution with mean zero and standard deviation 0.5, independent of $\mathbf{X}$. We consider a total of $2p+1$ subgroups, which were constructed as follows: $\mathcal{G}_0=\mathbb{R}^p$, $\mathcal{G}_k=\{\mathbf{X}:X_k>0,\mathbf{X}\in\mathbb{R}^p\}$, and $\mathcal{G}_{k+p}=\{\mathbf{X}:X_k\le 0,\mathbf{X}\in\mathbb{R}^p\}$ for $k=1,\ldots,p$. To obtain $\widehat{g}$, we set $s_2(\mathbf{X})=\mathbf{X}$ and applied linear regression with the adaptive lasso penalty, using $(\mathbf{X},T,\mathbf{X}T)$ as predictors. The regularization parameter was selected to minimize the mean squared error from the 10-fold cross-validation. Moreover, to determine variables in $\widehat{f}$, $\widetilde{f}^*$, and $\widetilde{f}$, we set $s(\mathbf{X})$ as the active variables in $\mathbf{X}$ in the above regression model, with the one-standard-error rule applied here for parsimony. We then applied the least squares method to $\{(s(\mathbf{X}_i),Y_i),i=1,\ldots,n\}$ to estimate $\widehat{f}$; similarly, we used $\{(s(\mathbf{X}_i),\widetilde{Y}_i),i=1,\ldots,n\}$ to estimate $\widetilde{f}$. On a bootstrapped data set, we applied the least squares method to $\{(s(\mathbf{X}_i^*),\widetilde{Y}_i^*),i=1,\ldots,n\}$ to obtain $\widetilde{f}^*$. The number of bootstrap samples was 500. The proposed method was compared to two simple alternatives:

- *LM*: Performs a linear regression analysis on the covariates $(T, \mathbf{X})$ and rejects the null hypothesis when T is significant (based on the one-sided Wald test at the nominal level);
- *LM2*: Performs a linear regression analysis on the covariates $(T, \mathbf{X}, T\mathbf{X})$ and rejects the null hypothesis based on an overall F test on the effects of $(T, T\mathbf{X})$.

We evaluated the proportion of rejection of different methods with the parameter a ranging from 0 to 1. In all of the simulations, the number of simulated datasets is 5000 when evaluating the type I error rates and is 2000 when evaluating the powers.

Note that LM is simply linear regression analysis focusing on the overall treatment effect, and LM2 uses the interaction terms to detect subgroup treatment effects. They are quick-and-dirty approaches for detecting the treatment effects. The proportions of rejection of the three methods are summarized in Table 1. The proposed test and the LM test approximately maintained the type I error rates in all the scenarios. When the true model was not a linear model, the LM test still yielded reasonably good type I error rates, while the LM2 test

TABLE 1
Simulation results in the absence of censoring

$n = 400$	I			II			III		
a	P	LM	LM2	P	LM	LM2	P	LM	LM2
0	0.049	0.051	0.054	0.049	0.051	0.054	0.052	0.056	0.050
0.25	0.312	0.316	0.086	0.350	0.321	0.169	0.149	0.158	0.088
0.5	0.816	0.731	0.265	0.890	0.729	0.734	0.481	0.328	0.303
0.75	0.992	0.940	0.635	0.999	0.936	0.995	0.851	0.557	0.710
1	1.000	0.992	0.905	1.000	0.990	1.000	0.980	0.759	0.962
	IV			V			VI		
a	P	LM	LM2	P	LM	LM2	P	LM	LM2
0	0.046	0.048	0.057	0.048	0.052	0.071	0.048	0.052	0.071
0.25	0.167	0.122	0.127	0.130	0.075	0.117	0.340	0.318	0.106
0.5	0.732	0.248	0.505	0.669	0.132	0.332	0.942	0.814	0.428
0.75	0.988	0.382	0.923	0.983	0.277	0.723	1.000	0.991	0.919
1	1.000	0.519	1.000	1.000	0.532	0.950	1.000	1.000	0.999
$n = 800$	I			II			III		
a	P	LM	LM2	P	LM	LM2	P	LM	LM2
0	0.047	0.047	0.053	0.047	0.047	0.053	0.047	0.051	0.046
0.25	0.562	0.506	0.165	0.618	0.509	0.424	0.262	0.230	0.160
0.5	0.994	0.953	0.690	0.996	0.952	0.997	0.804	0.569	0.747
0.75	1.000	0.999	0.982	1.000	0.998	1.000	0.996	0.838	0.999
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.957	1.000
	IV			V			VI		
a	P	LM	LM2	P	LM	LM2	P	LM	LM2
0	0.046	0.053	0.055	0.044	0.049	0.073	0.044	0.049	0.073
0.25	0.376	0.181	0.280	0.294	0.070	0.211	0.591	0.504	0.192
0.5	0.984	0.378	0.936	0.969	0.181	0.768	1.000	0.980	0.861
0.75	1.000	0.630	1.000	1.000	0.462	0.997	1.000	1.000	1.000
1	1.000	0.814	1.000	1.000	0.825	1.000	1.000	1.000	1.000

Note: The table reports the proportion of rejection for $n = 400$ and $n = 800$. "P" stands for the proposed omnibus test; "LM" and "LM2" stand for testing the effects of T and $(T, T\mathbf{X})$ under linear models, respectively.

yielded inflated type I error rates in scenarios V and VI. Compared to the LM and LM2 tests, the proposed omnibus test showed generally higher power in the scenarios we considered, sometimes very substantially.

In the second set of simulations, we considered censored outcomes, such as survival times. The survival times were generated as $\exp(Y)$, where Y was generated from scenarios (I)–(VI). The censoring time was generated from log-normal distributions, where the standard deviation of natural logarithm of the variable is fixed at 1 and the mean parameters were chosen to yield an approximate 25% rate of censoring. To estimate f and g, we first perform variable selection in $\mathbf{X}$ using censored rank independence screening proposed in Song et al. (2014). The treatment variable T was not used in the screening procedure. Estimation of $\widehat{f}$, $\widetilde{f}$, and $\widetilde{f}^*$ were based on a working AFT model $\log Y = \boldsymbol{\theta}^\top s(\mathbf{X}) + \epsilon$, where $s(\mathbf{X})$ includes the top five variables from marginal screening. To estimate g, we set $s_2(\mathbf{X}) = s(\mathbf{X})$ and used a working model $\log Y = \boldsymbol{\theta}^\top s_2(\mathbf{X}) + \gamma_0 T + \boldsymbol{\gamma}^\top s_2(\mathbf{X})T + \epsilon$. The fast-censored linear regression algorithm, proposed in Huang (2013), was applied to fit the AFT models. The proposed method was compared with commonly used models, including the Cox model and the AFT model. However, fitting AFT models with 50 covariates could be computationally unstable, because the weighted log-rank estimating functions are neither continuous nor, in general, monotone. Hence, we used the oracle tests where only active covariates, denoted by $\mathbf{X}_A$, are included in the AFT and the Cox models. More specifically, the comparisons used against our proposed test are:

- *Oracle AFT*: Performs an AFT model with covariates $(T, \mathbf{X}_A)$ and rejects the null hypothesis when T is significant (based on a one-sided test at the nominal level);
- *Oracle AFT2*: Performs an AFT model with covariates $(T, \mathbf{X}_A, T\mathbf{X}_A)$ and rejects the null hypothesis based on the Wald test of the effects on $(T, T\mathbf{X}_A)$;
- *Oracle Cox*: Performs a Cox model with covariates $(T, \mathbf{X}_A, T\mathbf{X}_A)$ and rejects the null hypothesis based on the Wald test of the effects on $(T, T\mathbf{X}_A)$.

The Oracle AFT and Oracle Cox tests cannot be carried out in practice but are chosen here as benchmarks for the analysis based on the AFT models and Cox models, respectively. The proportions of rejection of the four methods under comparison are summarized in Table 2. Due to model misspecification, the validity of variance estimates in Cox and the AFT models is not guaranteed, and thus the type I error rates may not approximate the nominal level. The proposed test and the Oracle AFT approach approximately maintained the type I error rates in all the scenarios. The Oracle AFT2 approach showed inflated type I errors in Scenarios IV, V, and VI; the Oracle Cox approach failed to maintain the appropriate type I error rates in all the scenarios. Our proposed test had power comparable to or higher than the Oracle AFT test.

In both sets of the simulation studies, we see that the proposed test maintained type I error rates reasonably well and had generally better power than their obvious alternatives that maintained the type I error rates, even when the alternative methods used oracle models for the covariates.

5. Real data analysis. We return to the randomized trial on panitumumab with FOLFIRI for treating mCRC. The regimen FOLFIRI has been considered as the standard chemotherapy for mCRC. Panitumumab is a monoclonal antibody directed against the EGFR and was approved in the United States as monotherapy for mCRC after disease progression with standard chemotherapy. In this study, patients were randomly assigned to two treatment arms: panitumumab-FOLFIRI and FOLFIRI only. The primary endpoint is the progression free survival, defined as the time to progression or death, whichever occurs first. The progression free survival time was subject to right censoring due to loss to follow-up and study end. The

TABLE 2
Simulation results in the presence of censoring

$n = 400$	I				II				III			
a	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O
0	0.051	0.161	0.050	0.061	0.051	0.165	0.050	0.064	0.056	0.117	0.051	0.063
0.25	0.236	0.311	0.231	0.206	0.205	0.530	0.158	0.390	0.191	0.343	0.165	0.229
0.5	0.616	0.744	0.583	0.664	0.532	0.959	0.308	0.934	0.419	0.817	0.372	0.734
0.75	0.855	0.954	0.850	0.940	0.863	1.000	0.518	1.000	0.741	0.992	0.572	0.979
1	0.963	0.996	0.965	0.996	0.985	1.000	0.713	1.000	0.936	1.000	0.743	1.000
	IV				V				VI			
a	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O
0	0.060	0.082	0.052	0.081	0.055	0.192	0.053	0.076	0.055	0.181	0.051	0.077
0.25	0.150	0.380	0.111	0.380	0.088	0.342	0.054	0.230	0.402	0.422	0.303	0.352
0.5	0.538	0.915	0.190	0.904	0.454	0.729	0.078	0.683	0.934	0.938	0.786	0.931
0.75	0.919	0.999	0.240	0.998	0.916	0.963	0.145	0.975	1.000	1.000	0.990	1.000
1	0.998	1.000	0.305	1.000	0.997	0.999	0.263	1.000	1.000	1.000	1.000	1.000
$n = 800$	I				II				III			
a	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O
0	0.050	0.154	0.047	0.053	0.050	0.160	0.045	0.051	0.055	0.116	0.050	0.053
0.25	0.392	0.472	0.368	0.350	0.337	0.781	0.237	0.672	0.258	0.555	0.254	0.419
0.5	0.894	0.942	0.818	0.919	0.886	1.000	0.534	1.000	0.720	0.987	0.568	0.974
0.75	0.996	0.999	0.984	1.000	0.999	1.000	0.793	1.000	0.976	1.000	0.827	1.000
1	1.000	1.000	0.999	1.000	1.000	1.000	0.920	1.000	1.000	1.000	0.953	1.000
	IV				V				VI			
a	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O	P	Cox-O	AFT-O	AFT2-O
0	0.053	0.084	0.055	0.081	0.055	0.196	0.049	0.067	0.055	0.191	0.048	0.077
0.25	0.253	0.614	0.144	0.613	0.185	0.489	0.061	0.404	0.658	0.637	0.474	0.596
0.5	0.906	0.996	0.278	0.995	0.856	0.934	0.116	0.957	1.000	1.000	0.973	0.999
0.75	1.000	1.000	0.406	1.000	1.000	1.000	0.204	1.000	1.000	1.000	1.000	1.000
1	1.000	1.000	0.479	1.000	1.000	1.000	0.419	1.000	1.000	1.000	1.000	1.000

Note: The table reports the proportion of rejection for $n = 400$ and $n = 800$. "P" stands for the proposed method; "AFT-O" and "AFT2-O" stand for the oracle AFT and oracle AFT2 approach, which are testing the effects of T and $(T, T\mathbf{X}_A)$ under AFT models using active covariates using Wald tests; "Cox-O" stands for testing the effects of $(T, T\mathbf{X}_A)$ using a Wald test.

covariates include tumor KRAS status (mutant vs. wild), sex (female vs. male), age (< 65 vs. ≥ 65), number of metastatic sites (< 3 vs. ≥ 3), and Eastern Cooperative Oncology Group (ECOG) performance status (0 vs. > 0). The ECOG performance status is a simple measure of functional status and determines ability of patient to tolerate therapies in serious illness. The threshold values on age, number of metastatic sites, and ECOG follows those in Peeters et al. (2010). KRAS mutations occur in approximately 35% to 43% of patients with mCRC. Previous studies have demonstrated patients with mCRC with mutant KRAS tumor status do not derive clinical benefit from anti-EGFR therapies. In our analysis we focus on 864 subjects who had complete covariate data and were under follow-up for at least one day. Among these patients, 431 of them were in the panitumumab-FOLFIRI arm ($T = 1$), and 433 were in the FOLFIRI arm ($T = 0$). Moreover, 463 subjects were in the KRAS wild subgroup; 514 subjects were males; 347 subjects were older than or equal to age 65; 453 subjects had more

TABLE 3
p-values of log rank test based on 11 *subgroups*

Group	p-value	Adjusted p-value
KRAS (wild)	0.011	0.123
Sex (M)	0.015	0.154
All	0.018	0.163
Age (< 65)	0.078	0.480
ECOG (> 0)	0.072	0.502
Metastatic sites (≥ 3)	0.090	0.541
Metastatic sites (< 3)	0.096	0.541
Age (≥ 65)	0.167	0.669
ECOG (0)	0.182	0.669
Sex (F)	0.553	1
KRAS (mutant)	0.681	1

Note: The adjusted p-values were obtained using Holm's method.

than two tumors. The median follow-up time was 168 days. In addition to the progression free survival, we also considered the response to treatment (yes/no) as a secondary endpoint.

We first applied the proposed test to the progression free survival. In this case, each of the five covariates was used to define subgroups. Each subgroup shares the same value on one of the covariates, resulting in five ways to partition the data and 10 subgroups. Adding in the entire population, we have a case of $J = 11$. The p-value of the proposed test, described in Section 3, was < 0.001. The subgroup with wild type KRAS status yielded the largest standardized difference.

We also consider partitioning the data finer by using the subgroups defined based on a pair of covariates. For example, the variable KRAS status and gender can be used to partition the data into eight possible subgroups: wild and male, wild and female, mutant and male, mutant and female, wild or male, wild or female, mutant or female, mutant or male. Adding the aforementioned 11 subgroups, we consider a total of $J = 91$ subgroups. The proposed test gave a p-value of 0.001. The subgroup of wild type KRAS status or male yielded the largest standardized difference.

The p-value of the log-rank test conducted in the wild KRAS group was 0.011, but it was no longer significant at the 0.05 level of significance if a multiplicity adjustment had been made for searching through the 11 subgroups (see Table 3). If we searched through the 91 subgroups, based on a single covariate or a pair of covariates, the p-value for the wild or male subgroup was 0.002 but, again, not significant enough to pass a multiplicity adjustment. The proposed omnibus test maintained desirable power even with many subgroup candidates considered in this case.

We also analyzed the binary response to treatment, as the secondary endpoint, using the test proposed in Section 2. The test using partitions, based on one and two covariates, yielded the p-value < 0.001, and the subgroup that yielded the largest standardized effect size was once again the wild type KRAS. Applying the univariate logistic regression analysis for the same data yielded p-values < 0.001 in the wild KRAS subgroup and 0.82 in the mutant KRAS subgroup. Therefore, for the secondary outcome, both the proposed omnibus test and the multiplicity adjustment (for $J = 11$) led to the conclusion that the treatment is effective in the wild KRAS group.

6. Discussion. The paper introduces a formal omnibus test on the existence of a subgroup with favorable treatment effects, where subgroups are constructed via partitions of the

covariate space. Such partitions are frequently employed in the search of subgroups in clinical trials, and the proposed test is a useful inferential tool to manage the risk of data snooping. While the treatment effects are often measured under a specific model, we note that the validity of the test relies little on model specification. The critical values of the proposed test are obtained via the bootstrap, and as a result, the proposed test is able to handle a large number of covariates/partitions and can accommodate various types of outcome variables. The effectiveness and robustness of the test are supported by its capability of preserving type-I error rates and good statistical power, as demonstrated by simulation studies and an application to clinical trial data.

When investigators look for promising results from many candidate subgroups, a common approach to controlling the type I error is through multiple test adjustments. However, candidate subgroups often have overlaps, so the subgroup-specific tests are not independent. As a result, commonly used multiplicity adjustments in p-values tend to result in loss of power. An important advantage of the proposed omnibus test is to avoid power loss due to conservatism in such multiplicity adjustments. When the test fails to detect subgroup effects, it suggests strongly not to pursue subgroups from the trial. If one wishes to keep looking for subgroups, multiplicity adjustments in the p-values need to be used.

The construction of the proposed test allows further interrogation of the specific subgroup constructions and identifying the most promising subgroups (e.g., associated with the largest standardized treatment effect estimates in the test statistic). However, fair quantification and formal inference about the treatment effect in the post hoc identified subgroups requires further investigation. We refer to Lee and Rubin (2016) and Guo and He (2021) for helpful exploration in that direction. To confirm the subgroup effects, it is often preferred to conduct an additional trial on a selected subgroup.

Acknowledgments. The authors would like to thank the referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Funding. Sun's research is partially supported by the National Institute of Health (NCI 5P30 CA013696 and NIA 2U19AG033655).

He's research is partially supported by NSF Award DMS-1914496.

Hu's research is partially supported by the National Institute of Health (NCI 5P30 CA013696, NIAID 1R01 AI143886, NIH/NCI 1R01 CA219896, NCI P01 CA098101).

SUPPLEMENTARY MATERIAL

Source code for "An omnibus test for detection of subgroup treatment effect via data partitioning" (Sun, He and Hu (2022)) (DOI: 10.1214/21-AOAS1589SUPP; .zip). An R package implementing the tests in this paper is also available online at https://github.com/yifeisun/subgroupTEtest. The data that support the findings of the panitumumab study in this paper are available from the Project Data Sphere's Data Sharing Platform. The general public can submit a request for downloading the data at the following URL: https://data.projectdatasphere.org/projectdatasphere/html/content/263.

REFERENCES

BEHR, M., ANSARI, M. A., MUNK, A. and HOLMES, C. (2020). Testing for dependence on tree structures. *Proc. Natl. Acad. Sci. USA* **117** 9787–9792. MR4236178 https://doi.org/10.1073/pnas.1912957117

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963 https://doi.org/10.1214/16-AOP1113

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2019). Inference on causal and structural parameters using many moment inequalities. *Rev. Econ. Stud.* **86** 1867–1900. MR4009488 https://doi.org/10.1093/restud/rdy065

DMITRIENKO, A., MILLEN, B. and LIPKOVICH, I. (2017). Multiplicity considerations in subgroup analysis. *Stat. Med.* **36** 4446–4454. MR3731226 https://doi.org/10.1002/sim.7416

DMITRIENKO, A., LIPKOVICH, I., DANE, A. and MUYSERS, C. (2020). Data-driven and confirmatory subgroup analysis in clinical trials. In *Design and Analysis of Subgroups with Biopharmaceutical Applications* (N. Ting, J. C. Cappelleri, S. Ho and D. Chen, eds.) 33–91. Springer, Berlin.

GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52** 203–223. MR0207130 https://doi.org/10.1093/biomet/52.1-2.203

GUO, X. and HE, X. (2021). Inference on selected subgroups in clinical trials. *J. Amer. Statist. Assoc.* **116** 1498–1506. MR4309288 https://doi.org/10.1080/01621459.2020.1740096

HUANG, Y. (2013). Fast censored linear regression. *Scand. J. Stat.* **40** 789–806. MR3145118 https://doi.org/10.1111/sjos.12031

JOSHI, N., FINE, J., CHU, R. and IVANOVA, A. (2019). Estimating the subgroup and testing for treatment effect in a post-hoc analysis of a clinical trial with a biomarker. *J. Biopharm. Statist.* **29** 685–695.

LEE, J. J. and RUBIN, D. B. (2016). Evaluating the validity of post-hoc subgroup inferences: A case study. *Amer. Statist.* **70** 39–46. MR3480669 https://doi.org/10.1080/00031305.2015.1093961

LOH, W.-Y. (2009). Improving the precision of classification trees. *Ann. Appl. Stat.* **3** 1710–1737. MR2752155 https://doi.org/10.1214/09-AOAS260

LOH, W.-Y., HE, X. and MAN, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Stat. Med.* **34** 1818–1833. MR3334694 https://doi.org/10.1002/sim.6454

LU, W., GOLDBERG, Y. and FINE, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99** 717–731. MR2966780 https://doi.org/10.1093/biomet/ass027

NAGGARA, O., RAYMOND, J., GUILBERT, F. and ALTMAN, D. G. (2011). The problem of subgroup analyses: An example from a trial on ruptured intracranial aneurysms. *Am. J. Neuroradiol.* **32** 633–636.

O'BRIEN, T. R., KUHS, K. A. and PFEIFFER, R. M. (2014). Subgroup differences in response to 8 weeks of ledipasvir/sofosbuvir for chronic hepatitis C. *Open Forum Infect. Dis.* **1** ofu110.

ONDRA, T., DMITRIENKO, A., FRIEDE, T., GRAF, A., MILLER, F., STALLARD, N. and POSCH, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *J. Biopharm. Statist.* **26** 99–119. https://doi.org/10.1080/10543406.2015.1092034

PEETERS, M., PRICE, T., CERVANTES, A., SOBRERO, A., DUCREUX, M., HOTKO, Y., ANDRÉ, T., CHAN, E., LORDICK, F. et al. (2010). Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. *Am. J. Clin. Oncol.* **28** 4706–4713.

PEETERS, M., OLINER, K. S., PRICE, T. J., CERVANTES, A., SOBRERO, A. F., DUCREUX, M., HOTKO, Y., ANDRÉ, T., CHAN, E. et al. (2015). Analysis of KRAS/NRAS mutations in a phase III study of panitumumab with FOLFIRI compared with FOLFIRI alone as second-line treatment for metastatic colorectal cancer. *Clin. Cancer Res.* **21** 5469–5479.

PEPE, M. S. and FLEMING, T. R. (1991). Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *J. Roy. Statist. Soc. Ser. B* **53** 341–352. MR1108331

SHEN, J. and HE, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110** 303–312. MR3338504 https://doi.org/10.1080/01621459.2014.894763

SHEN, J. and QU, A. (2020). Subgroup analysis based on structured mixed-effects models for longitudinal data. *J. Biopharm. Statist.* **30** 607–622.

SIES, A., DEMYTTENAERE, K. and MECHELEN, I. V. (2019). Studying treatment-effect heterogeneity in precision medicine through induced subgroups. *J. Biopharm. Statist.* **29** 491–507.

SONG, R., LU, W., MA, S. and JENG, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799–814. MR3286918 https://doi.org/10.1093/biomet/asu047

SUN, Y., HE, X. and HU, J. (2022). Supplement to "An omnibus test for detection of subgroup treatment effects via data partitioning." https://doi.org/10.1214/21-AOAS1589SUPP

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372. MR1041397 https://doi.org/10.1214/aos/1176347504

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735