Submitted to the Annals of Applied Statistics

SUBGROUP INFERENCE FOR MULTIPLE TREATMENTS AND MULTIPLE ENDPOINTS IN AN ALZHEIMER'S DISEASE TREATMENT TRIAL

By Patrick Schnell[1], Qi Tang[2], Peter Müller[3], and Bradley P. Carlin[1]

[1] *University of Minnesota;* [2] *AbbVie, Inc;* [3] *University of Texas*

Many new experimental treatments outperform the current standard only for a subset of the population. Subgroup identification methods provide estimates for the population subset which benefits most from treatment. However, when more than two treatments and multiple endpoints are under consideration, there are many possible requirements for a particular treatment to be beneficial. In this paper we adapt notions of decision-theoretic admissibility to the context of evaluating treatments in such trials. As an explicit demonstration of admissibility concepts we combine our approach with the method of credible subgroups, which in the case of a single outcome and treatment comparison provides Bayesian bounds on the benefiting subpopulation. We investigate our methods' performance via simulation, and apply them to a recent dataset from an Alzheimer's disease treatment trial. Our results account for multiplicity while showing patient covariate profiles that are (or are not) likely to be associated with treatment benefit, and are thus useful in their own right or as guide to patient enrollment in a second stage study.

1. Introduction. We develop a framework for subgroup analysis in clinical trials with more than two arms and multiple endpoints, and apply it to data from an Alzheimer's disease treatment trial. The proposed approach can be used together with any Bayesian subgroup analysis method that reports an estimate of the benefiting subpopulation, and generalizes the underlying method to many arms and multiple endpoints. Examples of estimates of the benefiting subpopulation include subsets which are likely contained in the benefiting subpopulation, and subsets for which the within-subset average treatment effect is positive. The approach becomes particularly meaningful if the underlying inference approach for subgroups includes posterior probability bounds for the reported subgroup.

Clinical trials have generally focused on demonstrating that an experimental treatment performs, on average, better than a control. Recently, re-

Keywords and phrases: Bayesian inference, Clinical trials, Heterogeneous treatment effect, Linear model, Simultaneous inference, Subgroup identification

searchers have become increasingly aware that the so-called "average treatment effect" (ATE) is problematic because it may not accurately represent the treatment effect for a particular patient. Treatments for complex diseases, such as cancer or Alzheimer's disease, may be effective for some patients but not for others. Some of this variation may be systematic heterogeneity in the treatment effect due to baseline characteristics such as age, genetic biomarkers, or disease progression. When investigators acknowledge treatment effect heterogeneity, their question is no longer "is there a treatment effect?", but rather, "who benefits from treatment?"

In the univariate case, the subgroup selection problem may be stated as follows: estimate the *benefiting subgroup*, i.e., a set of patients defined by a set of observable baseline characteristics for whom the personalized treatment effect (e.g. difference in expected outcome between treatments for that person) is positive. The traditional approach (e.g., Pocock et al. (2002); Dixon and Simon (1991)) is to test the average treatment effect in the entire population, perform hypothesis tests for treatment-covariate interactions, and then test subgroup-specific average treatment effects where an interaction is indicated. Well-known characteristics of this approach include insufficient power of interaction tests, lack of power for detecting subgroup-specific effects due to small subsample sizes, and multiplicity of hypothesis tests. However, the approach remains popular and attractive for its simplicity and parsimony.

Newer methods take different approaches, differing both in data models and modes of inference. Subgroup-based adaptive (SUBA) designs (Xu et al., 2016) adaptively allocate patients over the course of a trial to what is thought to be the best treatment, according to accumulated information and a recursive partitioning model, and report the final allocation scheme as its treatment recommendation for the broader population. Adaptive signature designs (Freidlin and Simon, 2005) and extensions (Freidlin, Jiang and Simon, 2010) construct via general classification models a subgroup thought to benefit and then perform a test for the average treatment effect in that subgroup. Extensive work has also been done in developing flexible models of treatment effects, especially those based on regression and classification trees (Breiman et al., 1984; Chipman, George and McCulloch, 1998, 2010).

Berger, Wang and Shen (2014) cast subgroup analysis as a model selection problem. They introduce a set of models arising from a tree splitting process for covariates that define potential subgroups, and use carefully chosen prior probability models. Posterior model probabilities formalize the desired subgroup determination, including posterior probabilities that a patient within a subgroup has a non-zero treatment effect and that the average within-subgroup treatment effect is non-zero. Sivaganesan, Laud and Müller (2011)

introduce an algorithm-based method by introducing decision boundaries on posterior model probabilities. Foster, Taylor and Ruberg (2011) describe subgroup analysis as inference for a subset B (A in their notation) in the covariate space that characterizes patients with substantially better treatment effect than average (or better than a pre-specified threshold). Here, treatment effects are defined as difference in mean response for two counterfactual outcomes under treatment and control in a two-arm trial. They then proceed to produce an estimate $\hat{\mathrm{B}}$ of B using tree-based methods, along with an estimate of the average treatment effect within $\hat{\mathrm{B}}$. Schnell et al. (2016) recognize B as an unknown quantity and proceed to characterize uncertainty by a "credible subgroup" pair (D, S) of sets such that $\mathrm{P}(\mathrm{D} \subseteq \mathrm{B} \subseteq \mathrm{S}|\text{data}) \geq 1-\alpha$. That is, D and S are subsets in the covariate space with a bound on the posterior probability that these two sets bracket the desired benefiting subset B. In the context of this setup, some of the earlier discussed methods can be described as reporting a set similar to D (though not necessarily with the same bounding relationship to B), for example when selecting covariates and thresholds that define a subset D with substantially higher within-subgroup average treatment effect than the overall treatment effect. However, Schnell et al. (2016) appears to be the first to consider the notion of reporting an encompassing set S.

In this paper, we develop subgroup analysis methods to handle cases in which *more than two treatments* are being compared with respect to *multiple endpoints*. This multivariate problem setting admits several ways of defining a treatment effect and benefiting subgroup, as well as strategies for choosing the multiplicities for which to adjust. The initial discussion is general and can be applied with any method that reports (or from which can be extracted) an estimate for a benefiting subgroup B, especially when inference includes posterior probabilities $\mathrm{P}(\mathrm{D} \subseteq \mathrm{B}|\text{data})$. Here D is the reported estimate, together with the posterior probability that the reported subset does indeed characterize covariate combinations with a substantially higher treatment effect (or one exceeding some other threshold). The discussion can also apply to methods which produce some subgroup by any means and then tests for a within-subgroup treatment effect, though these methods will not be our focus here. Eventually, in the implementation we will incorporate the approach of Schnell et al. (2016), who report a credible subgroup pair (D, S). Additionally, in that case, we offer extended machinery for performing the multiple testing procedure for arbitrary joint posterior distributions of personalized treatment effects, which allows the use of a much wider variety of error distributions and regression models than the normal-errors linear regression to which Schnell et al. (2016) is restricted.

A related course of research is underway in the area of dynamic treatment regimes (DTRs), which infers optimal processes in which sequences of treatments are given to a single patient in a response-adaptive manner. Several methods have been developed to select the best from among many previously vetted treatments for individual patients in the presence of multiple relevant endpoints. Thall, Sung and Estey (2002) treat response, non-response, and death as an ordinal outcome and use a real-valued utility function elicited from experts to quantify the trade-off between response and death. Thall et al. (2007) report a trial in which four treatments were tested in a two-stage regime. Almirall, Lizotte and Murphy (2012) include patient preference among various endpoints in the estimated rule, in addition to clinical characteristics. Lizotte, Bowling and Murphy (2012) identify optimal treatment regimes for all linear combinations of endpoints, while Laber, Lizotte and Ferguson (2014) and Lizotte and Laber (2016) report regimes with sets of non-inferior treatment choices. Since research in DTRs focuses on providing optimal care to a given patient, attention is not generally paid to Type I error control. In contrast, our work focuses on single-stage, population-level inferences for a given treatment, and owing to our focus on the regulatory process, attention must be paid to Type I error and its control under multiplicity of endpoints, treatments, and covariate profiles.

Our motivating data set stems from a clinical trial of an Alzheimer's disease (AD) treatment carried out by AbbVie. While effective treatment strategies for AD are in their infancies, a number of risk factors for the disease are known. For example, advanced age and the presence of the ApoE4 allele dramatically increase the risk of AD, while longer education and higher intelligence appear somewhat protective (Burns and Iliffe, 2009). It is possible that some of these prognostic factors are also predictive of the effectiveness of certain treatments. In this motivating trial, we are interested in both the efficacy and safety of the test treatment relative to two controls (placebo and active), and wish to allow for heterogeneity in those effects. Our primary interest lies in searching for benefiting subgroups in a study that, like most, failed to show an overall treatment benefit in the population as a whole, but which may still reveal important subgroups, either as candidates for immediate approval or to be further studied in subsequent trials. This is very valuable, since if a strong effect is identified in a subgroup, it is sometimes possible, even post hoc, to file a new drug application in settings where the disease is severe and there is no effective standard of care.

The remainder of our paper is organized as follows. Section 2 develops an inference framework for trials with more than two arms and multiple endpoints, with Section 2.4 reviewing and extending the concept of credible

subgroups in this setting. Section 3 applies the methods of Section 2; in particular, Section 3.1 provides simulation results regarding the methods' sensitivity, specificity, and Type I error, while Section 3.2 illustrates the use of a subset of the methods on the motivating data set with two endpoints and five arms. Finally, Section 4 discusses our findings and offers avenues for further research.

2. Inference on Subgroups for Multiple Endpoints and Many Arms.

2.1. *Notation.* Consider a patient population represented by a covariate space C; for example, in Section 3.2 we consider a trial with Alzheimer's disease patients between the ages of 55 and 90, with covariates specifying age, sex, disease severity, and carrier status of a genetic biomarker. When investigating a treatment, it is desirable to make inferences regarding the subset of this population which benefits from the treatment over the control. In particular, we define the *benefiting subgroup* B as the set of covariate vectors $\boldsymbol{z}$ for which the *personalized treatment effect* $\Delta(\boldsymbol{z})$ is greater than some fixed threshold of clinical significance δ. The treatment effect may be, for example, a difference in expected response, a log odds ratio, or a log hazard ratio, signed so that a positive treatment effect indicates benefit. Except for special model assumptions, like the proportional hazards rate model, the use of log odds ratios would include a suitable definition of averaging over other event times and other covariates. Where $\Delta(\boldsymbol{z}) > \delta$, we simply say that the treatment is *beneficial at* $\boldsymbol{z}$. We begin by considering a simple estimator D for B, constructed so that for each $\boldsymbol{z} \in \mathrm{D}$, $\mathrm{P}(\boldsymbol{z} \in \mathrm{B}|\mathrm{data}) \geq 1-\alpha$ and later extend via multiplicity adjustments to estimators such that $\mathrm{P}(\mathrm{D} \subseteq \mathrm{B}|\mathrm{data}) \geq 1-\alpha$ or $\mathrm{P}(\mathrm{D} \subseteq \mathrm{B} \subseteq \mathrm{S}|\mathrm{data}) \geq 1-\alpha$.

2.2. *Subgroup Inference for Multiple Endpoints.* Results regarding the effect of a treatment on a specific endpoint are generally not considered in a vacuum. For example, an experimental treatment may have approximately the same effect as the standard of care on the primary endpoint (cognitive function score in our example), but have a lower instance of adverse side effects such as nausea. In such a situation, it would be useful to know not only who benefits from the experimental treatment with respect to the primary endpoint, but also who is likely to avoid side effects.

Suppose that there are $K \geq 2$ endpoints by which the test treatment is being compared to the control, and let $\Delta_k(\boldsymbol{z})$ be the treatment effect at covariate point $\boldsymbol{z}$ with respect to the kth endpoint. It is possible to construct subgroup inferences for the treatment effect corresponding to each endpoint,

either independently or adjusting for the multiplicity of endpoint inferences. For a set of independently estimated subgroups $\{\mathrm{D}_k\}_{k=1}^K$, we have that for each endpoint k and covariate point $\boldsymbol{z} \in \mathrm{D}$, $\mathrm{P}(\boldsymbol{z} \in \mathrm{B}_k|\text{data}) \geq 1-\alpha$. A set of subgroups is *simultaneous* (adjusting for endpoint multiplicity) if $\mathrm{P}(\{k : \boldsymbol{z} \in \mathrm{D}_k\} \subseteq \{k : \boldsymbol{z} \in \mathrm{B}_k\}|\text{data}) \geq 1-\alpha$ for each $\boldsymbol{z}$. Both methods result in K subgroup estimates, and may be used when each of the endpoints are of interest separately, rather than in combination.

A way to construct a single subgroup estimate that incorporates information about each of the endpoint effects is through a *utility function*, e.g., trading off probability of response and risk of death as in Thall, Sung and Estey (2002). Let u be some utility function of all the endpoints, and define the treatment effect $\Delta_u(\boldsymbol{z})$ as $\mathrm{E}[u|\boldsymbol{z}, t=1] - \mathrm{E}[u|\boldsymbol{z}, t=0]$, where $t=1$ indicates the test treatment and $t=0$ the control. The benefiting subset B and the subgroup estimate D may then be defined in the same way as in the single-endpoint case. Constructing a single subgroup estimate may simplify interpretation, but it is often difficult for multiple parties to agree on a single, often stylized utility function, especially for diseases such as Alzheimer's that affect quality of life in complex ways and frequently have uncomfortable side effects. If a range or distribution U of utility functions is to be considered, $\Delta_U(\boldsymbol{z})$ may be constructed to reflect some summary of the distribution of the $\Delta_u(\boldsymbol{x})$ as u varies, such as the mean, median, or minimum.

We can also construct a joint subgroup report motivated by the decision-theoretic concept of *admissibility*. Recall that a decision rule is admissible if there are no other rules that always perform at least as well and better in at least one case. Here we would like to call a test treatment *admissible at* $\boldsymbol{z}$ if the control treatment does not perform at least as well with respect to *every* endpoint and better with respect to at least one endpoint for a patient with covariate vector $\boldsymbol{z}$. Strictly speaking, the formalization of this definition is that a treatment is admissible at $\boldsymbol{z}$ unless $\Delta_k(\boldsymbol{z}) \leq 0$ for all k and the inequality is strict for at least one k.

Next, we generalize to allow for thresholds of clinical significance and noninferiority. In addition to the δ_k, the thresholds for clinical significance, let $\varepsilon_k \leq \delta_k$ be thresholds for non-inferiority, i.e., a treatment is considered "just as good" if $\varepsilon_k \leq \Delta_k(\boldsymbol{z}) \leq \delta_k$. Introducing these thresholds allows for multiple formulations of criteria. We call a treatment *weakly admissible at* $\boldsymbol{z}$ if $\Delta_k(\boldsymbol{z}) > \delta_k$ for at least one k *or* $\Delta_k(\boldsymbol{z}) \geq \varepsilon_k$ for all k. This is the generalization of our previous definition of admissibility most directly related to the decision-theoretic concept. However, a treatment may be undesirable if it is demonstrably inferior with respect to one endpoint, even if it is

superior in others, or if it is not superior in any. Thus we call a treatment *strongly admissible at* $\boldsymbol{z}$ if $\Delta_k(\boldsymbol{z}) > \delta_k$ for at least one k *and* $\Delta_k(\boldsymbol{z}) \geq \varepsilon_k$ for all k. A related method is to require only *noninferiority at* $\boldsymbol{z}$, i.e., that $\Delta_k(\boldsymbol{z}) \geq \varepsilon_k$ for all k.

The decision-theoretic criteria described above may be written in notation unified with the previous formulations of individual-endpoint and utility function treatment effects. For example, we define $\mathbb{I}$(condition) to be 1 if the condition is true and 0 otherwise, an indicator of strong admissibility may be written as

$$\Delta_{sa}(\boldsymbol{z}) = \mathbb{I}\left[\max_k \{\Delta_k(\boldsymbol{z}) - \delta_k\} > 0\right] \mathbb{I}\left[\min_k \{\Delta_k(\boldsymbol{z}) - \varepsilon_k\} \geq 0\right] \tag{1}$$

and compared to $\delta_{sa} = 0$ (with sa indicating strong admissibility) in the same fashion as the treatment effects above. We may similarly define the indicator $\Delta_{wa}(\boldsymbol{z})$ for weak admissibility (wa), which would then be compared to $\delta_{wa} = 0$. We can then define B as the set of $\boldsymbol{z}$ for which the treatment is admissible, and construct the desired joint subgroup report in the usual fashion. We term this approach the *direct method* for estimating admissibility.

A multiplicity problem arises when constructing subgroup reports from Δ_{sa} or Δ_{wa}. As more endpoints are included in an analysis, the frequentist probability of identifying at least one endpoint with respect to which the test treatment is superior or inferior increases, even when treatments are equivalent with respect to every endpoint. This makes it more likely for a treatment to be classified as weakly admissible or not strongly admissible. To avoid these biases, we may construct admissibility inferences via a *fully adjusted method* as follows. Let $\{\mathrm{D}_k\}_{k=1}^K$ be a simultaneous set of subgroup reports for superiority with respect to the K endpoints such that for all $\boldsymbol{z}$, $\mathrm{P}(\{k : \boldsymbol{z} \in \mathrm{D}_k\} \subseteq \{k : \boldsymbol{z} \in \mathrm{B}_k\}|\text{data}) \geq 1 - \alpha$, and $\{\mathrm{D}'_k\}_{k=1}^K$ be similarly defined for non-inferiority. Then for weak and strong admissibility, respectively,

$$\mathrm{D}_{wa} = \left\{\bigcup_{k=1}^K \mathrm{D}_k\right\} \cup \left\{\bigcap_{k=1}^K \mathrm{D}'_k\right\}, \quad \mathrm{D}_{sa} = \left\{\bigcup_{k=1}^K \mathrm{D}_k\right\} \cap \left\{\bigcap_{k=1}^K \mathrm{D}'_k\right\}. \tag{2}$$

2.3. *Many-Arm Multiple-Endpoint Subgroup Inferences.* Suppose now that there are $M > 2$ treatments being considered. It may not be desired to compare every treatment to every other. For example, we may envision a scenario in which there are three test treatments and one control, and it is desired to determine for each test treatment which patients benefit relative to the

control. Consider a *competition graph* $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{t = 1, \ldots, M\}$ is the set of treatment arm vertices and $\mathcal{E} = \{(t, c)\}$ is the set of directed edges where (t, c) is present if treatment t is being compared to control c. Let $\mathcal{E}(t)$ be the set of edges which originate at t. Let $\Delta_k^{tc}(\boldsymbol{z})$ be the effect of treatment t relative to treatment c for endpoint k, and δ_k^{tc} be a threshold of clinical significance such that $\Delta_k^{ct}(\boldsymbol{z}) = -\Delta_k^{tc}(\boldsymbol{z})$ but δ_k^{ct} is not necessarily the same as δ_k^{tc}. We generalize each of the two-arm methods to the many-arm multiple-endpoint case.

A subgroup inference may be constructed for each of the $K|\mathcal{E}|$ endpoint-comparison combinations, either independently or simultaneously (adjusting for multiplicity among endpoints and comparisons). For a set of independently generated inferences $\{\mathrm{D}_k^{tc}\}$, we have that for each endpoint-comparison pair $(k, (t, c))$ and covariate point $\boldsymbol{z} \in \mathrm{D}_k^{tc}$, $\mathrm{P}(\boldsymbol{z} \in \mathrm{B}_k^{tc}|\text{data}) \geq 1 - \alpha$. For a simultaneous set of inferences we require that for each $\boldsymbol{z}$, $\mathrm{P}(\{(k, (t, c)) : \boldsymbol{z} \in \mathrm{D}_k^{tc}\} \subseteq \{(k, (t, c)) : \boldsymbol{z} \in \mathrm{B}_k^{tc}\}|\text{data}) \geq 1 - \alpha$, where (t, c) varies over $\mathcal{E}$. These methods may be useful when each of the endpoints and treatment-comparisons are of interest separately.

Alternatively, inferences may be constructed for each of the KM endpoint-treatment combinations, in which each treatment t is compared against the totality of its competition, the comparison being denoted as $t*$. Again, the estimates D_k^{t*} may be determined independently or simultaneously. Let $\Delta_k^{t*}(\boldsymbol{x}) = \min_{c \in \mathcal{E}(t)} \left\{\Delta_k^{tc}(\boldsymbol{z}) - \delta_k^{tc}\right\}$ be the treatment effect versus the totality of competition and $\delta_k^{t*} = 0$ be the corresponding threshold, so that t is considered beneficial if it outperforms all of its competition by the corresponding margins. For independent sets of pairs, we require that for each (k, t) and $\boldsymbol{z} \in \mathrm{D}_k^{t*}$, $\mathrm{P}(\boldsymbol{z} \in \mathrm{B}_k^{t*}|\text{data}) \geq 1 - \alpha$. For a simultaneous set of inferences, we would require for each $\boldsymbol{z}$, $\mathrm{P}(\{(k, t) : \boldsymbol{z} \in \mathrm{D}_k^{t*}\} \subseteq \{(k, t) : \boldsymbol{z} \in \mathrm{B}_k^{t*}\}|\text{data}) \geq 1 - \alpha$.

Utility functions may be used to reduce the effective number of endpoints to one, and either $|\mathcal{E}|$ inferences may be constructed for pairwise treatment effects Δ_u^{tc}, or M may be constructed for the treatment effects Δ_u^{t*}. Alternatively, inferences for weak and strong admissibility or noninferiority may be constructed, either for a treatment against each of its competitors separately (e.g. with respect to each Δ_{sa}^{tc}), or for a treatment against the totality of its competition (e.g. with respect to Δ_{sa}^{t*}). Again, sets of credible subgroup pairs may be constructed independently or simultaneously. If using admissibility inferences corrected for multiplicity as in (2), a similar multiplicity adjustment may be made for many arms by taking, for weak and strong

admissibility, respectively,

$$\mathrm{D}_{wa}^{t*} = \bigcap_{(t,c)\in\mathcal{E}(t)} \mathrm{D}_{wa}^{tc}, \quad \mathrm{D}_{sa}^{t*} = \bigcap_{(t,c)\in\mathcal{E}(t)} \mathrm{D}_{sa}^{tc}. \tag{3}$$

2.4. *Implementation via Credible Subgroups.* We now develop in detail the implementation of the general approach for the adjustment for multiple endpoints and multiple treatment comparisons when the underlying model is the report of credible subgroup pairs as proposed in Schnell et al. (2016). This implementation is particularly interesting because it simplifies the form of certain probability statements by adjusting for multiplicity not only of endpoints and treatments, but covariate points as well.

The function of a *credible subgroup pair* is to attempt to bound the benefiting subgroup by two credible subgroups, one that is contained in the benefiting subgroup and one that contains it. Formally, an *exclusive credible subgroup* D and an *inclusive credible subgroup* S constitute a credible subgroup pair (D, S) if the posterior probability that $\mathrm{D} \subseteq \mathrm{B} \subseteq \mathrm{S}$ is at least $1-\alpha$, i.e. $\mathrm{P}(\mathrm{D} \subseteq \mathrm{B} \subseteq \mathrm{S}|\text{data}) \geq 1-\alpha$. This is analogous to the definition of credible intervals in one-dimensional space: $\mathrm{P}(L \leq \theta \leq U|\text{data}) \geq 1-\alpha$. The general procedure for constructing the credible subgroup pair (D, S) in the univariate case is to perform a regression of the personalized treatment effect $\Delta(\boldsymbol{z})$ on the predictive covariates $\boldsymbol{z}$, construct simultaneous credible bounds for the regression surface, and take as D points where the lower bound exceeds the threshold δ and as S those where the upper bound exceeds δ. When considering multiple endpoints and many treatments, the probability statements satisfied by the construction are $\mathrm{P}(\mathrm{D}_k^{tc} \subseteq \mathrm{B}_k^{tc} \subseteq \mathrm{S}_k^{tc}|\text{data}) \geq 1-\alpha$ for independent pairs, and $\mathrm{P}(\forall(k,(t,c)) \in \{1,\ldots,K\}\times\mathcal{E},\ \mathrm{D}_k^{tc} \subseteq \mathrm{B}_k^{tc} \subseteq \mathrm{S}_k^{tc}|\text{data}) \geq 1-\alpha$ for simultaneous pair sets.

We illustrate the general derivation of simultaneous sets of credible subgroups for treatment effects of the form $\Delta_k^{tc}(\boldsymbol{z})$ on a restriction C of the covariate space. This is the most general class presented, whence other classes discussed previously can be seen as special cases.

A simultaneous set of credible subgroup pairs is derived from the joint distribution of many treatment effects corresponding to various covariate points, endpoints, and treatment comparisons. Let $\widehat{\Delta}_k^{tc}(\boldsymbol{z}) = \mathrm{E}\left[\Delta_k^{tc}(\boldsymbol{z}) \mid \text{data}\right]$. Simultaneous credible bands for the $\Delta_k^{tc}(\boldsymbol{z})$ on C may be constructed, by an extension of Uusipaikka (1983), as

$$\Delta_k^{tc}(\boldsymbol{z}) \in \widehat{\Delta}_k^{tc}(\boldsymbol{z}) \pm \sqrt{W_\alpha^* \,\mathrm{Var}[\Delta_k^{tc}(\boldsymbol{z})]} \tag{4}$$

where W_α^* is the $1-\alpha$ quantile of the distribution of

$$W = \sup_{(\boldsymbol{z},k,(t,c))} \frac{\{\Delta_k^{tc}(\boldsymbol{z}) - \widehat{\Delta}_k^{tc}(\boldsymbol{z})\}^2}{\mathrm{Var}[\Delta_k^{tc}(\boldsymbol{z})]}. \tag{5}$$

and $(\boldsymbol{z}, k, (t,c))$ ranges over $\mathrm{C} \times \{1,\ldots,K\} \times \mathcal{V}$. The value of W_α^* may be estimated from a sample from the joint posterior of the $\Delta_k^{tc}(\boldsymbol{z})$.

The use of (4) is most appropriate when the posterior distributions of the $\Delta_k^{tc}(\boldsymbol{z})$ are continuous and differ only by a scale parameter. When discontinuous posterior distributions are present, for instance that of $\Delta_{sa}(\boldsymbol{z})$ in (1), a quantile-based credible band may be more appropriate. Let $F(y) = \mathrm{P}[Y \le y]$, $F^{-1}(p) = \inf\{y:\ p \le F(y)\}$, $G(y) = \mathrm{P}[Y < y]$, and $G^{-1}(p) = \sup\{y:\ p \ge G(y)\}$. If W_α^* is the α quantile of the distribution of

$$W = \inf_{(\boldsymbol{z},k,(t,c))} \min\left\{F_{\Delta_k^{tc}(\boldsymbol{z})}\left[\Delta_k^{tc}(\boldsymbol{z})\right],\ 1 - G_{\Delta_k^{tc}(\boldsymbol{z})}\left[\Delta_k^{tc}(\boldsymbol{z})\right]\right\}, \tag{6}$$

then

$$\Delta_k^{tc}(\boldsymbol{z}) \in \left[F^{-1}_{\Delta_k^{tc}(\boldsymbol{z})}\left(W_\alpha^*\right),\ G^{-1}_{\Delta_k^{tc}(\boldsymbol{z})}\left(1 - W_\alpha^*\right)\right] \tag{7}$$

is a $1-\alpha$ simultaneous credible band (Schnell et al. (2017), Theorem 3). Distribution functions and W_α^* may be estimated from a sample from the joint posterior of the $\Delta_k^{tc}(\boldsymbol{z})$.

Given simultaneous credible bands such as those in (4) and (7), the exclusive credible subgroups D_k^{tc} and inclusive credible subgroups S_k^{tc} are constructed by comparing the upper and lower bounds of the bands to δ_k^{tc}. In the case of (4), the exclusive credible subgroup D_k^{tc} and inclusive credible subgroup S_k^{tc} are given by

$$\begin{aligned}
\mathrm{D}_k^{tc} &= \left\{\boldsymbol{z} \in \mathrm{C}:\ \widehat{\Delta}_k^{tc}(\boldsymbol{z}) - \sqrt{W_\alpha^* \,\mathrm{Var}\left[\Delta_k^{tc}(\boldsymbol{z})\right]} > \delta_k^{tc}\right\}, \\
\mathrm{S}_k^{tc} &= \left\{\boldsymbol{z} \in \mathrm{C}:\ \widehat{\Delta}_k^{tc}(\boldsymbol{z}) + \sqrt{W_\alpha^* \,\mathrm{Var}\left[\Delta_k^{tc}(\boldsymbol{z})\right]} \ge \delta_k^{tc}\right\},
\end{aligned} \tag{8}$$

and $\mathrm{P}(\mathrm{D}_k^{tc} \subseteq \mathrm{B}_k^{tc} \subseteq \mathrm{S}_k^{tc} | \mathrm{data}) \ge 1-\alpha$. The loose inequality is used for S_k^{tc} so that if $\delta_k^{tc} = 0 = \delta_k^{ct}$ then $\mathrm{D}_k^{ct} = \left(\mathrm{S}_k^{tc}\right)^{\mathsf{c}}$. Credible subgroups derived from the form (7) are constructed similarly.

Once the $(\mathrm{D}_k^{tc}, \mathrm{S}_k^{tc})$ are available, credible subgroups for admissibility may be constructed through the following analogs of equations (2) and (3):

$$(\mathrm{D}_{wa}, \mathrm{S}_{wa}) = \left(\left\{\bigcup_{k=1}^K \mathrm{D}_k\right\} \cup \left\{\bigcap_{k=1}^K \mathrm{D}'_k\right\},\quad \left\{\bigcup_{k=1}^K \mathrm{S}_k\right\} \cup \left\{\bigcap_{k=1}^K \mathrm{S}'_k\right\}\right), \tag{9}$$

$$(10)\quad (\mathrm{D}_{sa}, \mathrm{S}_{sa}) = \left(\left\{ \bigcup_{k=1}^{K} \mathrm{D}_k \right\} \cap \left\{ \bigcap_{k=1}^{K} \mathrm{D}'_k \right\}, \quad \left\{ \bigcup_{k=1}^{K} \mathrm{S}_k \right\} \cap \left\{ \bigcap_{k=1}^{K} \mathrm{S}'_k \right\} \right).$$

$$(11)\quad \begin{aligned} \left(\mathrm{D}^{t*}_{wa}, \mathrm{S}^{t*}_{wa}\right) &= \left(\bigcap_{(t,c)\in\mathcal{E}(t)} \mathrm{D}^{tc}_{wa}, \bigcup_{(t,c)\in\mathcal{E}(t)} \mathrm{S}^{tc}_{wa} \right), \\ \left(\mathrm{D}^{t*}_{sa}, \mathrm{S}^{t*}_{sa}\right) &= \left(\bigcap_{(t,c)\in\mathcal{E}(t)} \mathrm{D}^{tc}_{sa}, \bigcup_{(t,c)\in\mathcal{E}(t)} \mathrm{S}^{tc}_{sa} \right). \end{aligned}$$

3. Results.

3.1. *Simulations.* We perform a simulation study to evaluate certain frequentist properties of each method for finding credible subgroup pairs. We are primarily concerned with the properties of our four different types of admissibility: weak and strong, each estimated via the fully adjusted and direct methods. Our operating characteristics of primary interest are the average sensitivity and specificity of the exclusive credible subgroup D under increasing numbers of endpoints and treatment arms.

Each simulated data set is produced with A arms, $N = 100$ patients per arm, K endpoints, and $P = 3$ covariates. For patient i in arm a, $\boldsymbol{x}_{ai} = (1, x_{ai2}, x_{ai3})$ is a prognostic covariate vector where x_{ai2} and x_{ai3} are discrete covariates randomly drawn from $\{-2, -1, 0, 1, 2\}$ with probabilities $\{1/16, 1/4, 3/8, 1/4, 1/16\}$, respectively. The same vector is used as the predictive covariate vector: $\boldsymbol{z}_{ai} = \boldsymbol{x}_{ai}$. The following model is used to produce the simulated data:

$$(12)\qquad Y_{aik} | \eta_{aik}, \sigma^2_k \sim \mathrm{Normal}(\eta_{aik}, \sigma^2_k), \quad \eta_{aik} \quad = \boldsymbol{x}'_{ai}\boldsymbol{\beta}_k + \boldsymbol{z}'_{ai}\boldsymbol{\gamma}^{(a)}_k,$$

where Y_{aik} is the response in the kth endpoint for patient i in arm a, $\boldsymbol{\beta}_k \equiv (1, 1, 1)$ for all k, and the $\boldsymbol{\gamma}^{(a)}_k$ are determined as follows: $\boldsymbol{\gamma}^{(a)}_1 = (0, 1/3, 0)$ for $1 < a < A$, $\boldsymbol{\gamma}^{(A)}_1 = (0, 1, 0)$, all other $\boldsymbol{\gamma}^{(a)}_k = (0, 0, 0)$. The scenarios tested were $A = 2 - 8$ with $K = 1$, and $K = 1 - 8$ with $A = 2$. We simulated 1000 data sets per scenario, constructing 50% credible subgroup pairs.

The model used to fit the simulated data is the same, with priors $\sigma^2_k \sim \mathrm{InverseGamma}(10^{-4}, 10^{-4})$, $\beta_{kp} \sim \mathrm{Normal}(0, 10^4)$ for all k, p, and $\gamma^{(1)}_{kp} = 0$, $\gamma^{(a)}_{k1} \sim \mathrm{Normal}(0, 10^4)$, $\gamma^{(a)}_{kp} \sim \mathrm{Normal}(0, 1)$ for $a > 1$ and all k, p. The model was fit using the `NIMBLE` `R` package (NIMBLE Development Team, 2015) for 100 burn-in iterations and an additional 1000 recorded iterations for each simulated data set. Credible subgroups were constructed using (7).

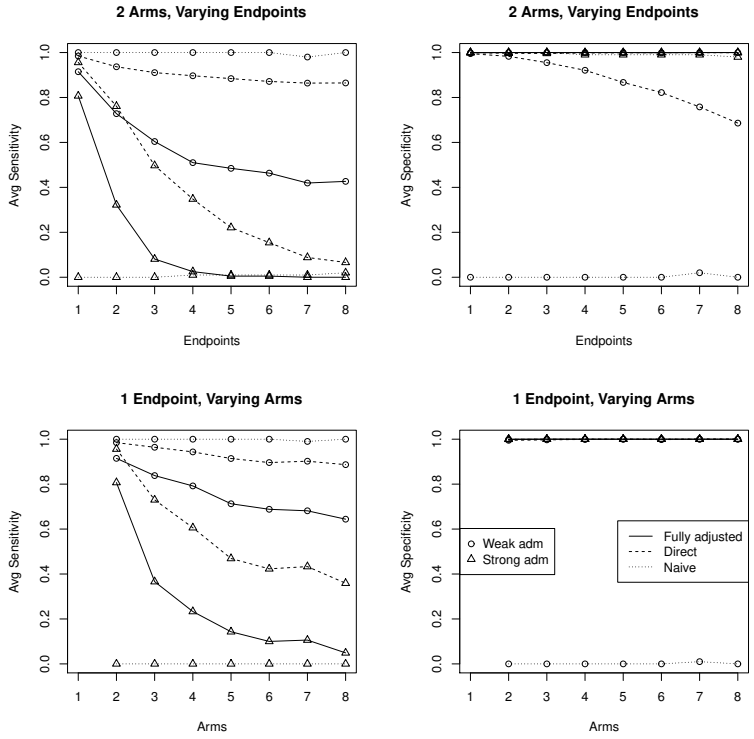

FIG 1. *Simulated sensitivity (left column) and specificity (right column) for a study with $A = 2$ arms and varying number of endpoints (top row) and a study with $K = 1$ endpoints and a varying number of arms (bottom row). In most cases, sensitivity falls and specificity remains high as more arms or endpoints are added.*

We also compare the fully adjusted and direct methods to a "naive" method for determining admissibilities. We use the above regression model without treatment-covariate interactions to estimate an average treatment effect independently for each endpoint-treatment combination. For each draw from the joint posterior of the average treatment effects we compute draws of weak and strong admissibility, then use the posteriors of the admissibilities to make inferences at the 50% level. The direct method reduces to the naive method when there are no treatment-covariate interactions.

The results of the simulation study are displayed in Figure 1. In most cases, sensitivity falls and specificity remains high as the number of arms or endpoints increases, with the exception that the specificity of direct weak admissibility decreases as endpoints are added. Additionally, detection of strong admissibility is more difficult than detection of weak admissibility,

and adjusting for multiplicity in the estimation of admissibilities (i.e. using the fully adjusted instead of direct method) reduces sensitivity. The naive approach retains very high sensitivity and very low specificity for weak admissibility in all presented scenarios, and very low sensitivity and very high specificity for strong admissibility in all presented scenarios.

3.2. *Application to Alzheimer's Disease Data Set.* We illustrate the extended credible subgroups methods on a data set derived from a clinical trial of an Alzheimer's disease treatment explored by AbbVie. Three doses (low, medium, high) of an experimental treatment are to be compared to active control and to a placebo. Baseline measurements for disease severity, age, sex, and carrier status of a genetic biomarker constitute covariates. After 24 weeks of treatment, two endpoints are of interest: improvement (negative change in disease severity) as the efficacy endpoint, and the reporting of at least one adverse event indicated by the attending physician to be possibly related to the treatment. We consider only those patients who have complete observations, which yields a data set of 331 patients. All covariates and the efficacy outcome are standardized for the analysis and displayed in their original units.

Let $a = 0, 1, 2, 3, 4$ denote the placebo, low, medium, and high doses of the test treatment, and active control treatment arms, respectively. For patient i, let Y_{ik} for $k = 1, 2$ denote the change in severity (continuous) and adverse event occurrence (binary) endpoints, respectively, and $\boldsymbol{x}_{i1} = \boldsymbol{x}_{i2} = \boldsymbol{z}_{i1} = \boldsymbol{z}_{i2}$ be the prognostic and predictive covariate vectors for each endpoint (including intercept, all considered as equal here). Let $\boldsymbol{\beta}_k$ be the vector of prognostic effects for the kth endpoint, and $\boldsymbol{\gamma}_k^{(a)}$ be the vector of predictive effects for the kth endpoint and treatment arm a, with $\boldsymbol{\gamma}_k^{(0)} = \mathbf{0}$. Also let $d^{(a)}$ be a scalar representing the level of activity of the drug dose in arm a compared to the maximum dose of the same drug, with $0 = d^{(0)} \leq d^{(1)} \leq d^{(2)} \leq d^{(3)} = d^{(4)} = 1$ and $\boldsymbol{\gamma}_k^{(1)} = \boldsymbol{\gamma}_k^{(2)} = \boldsymbol{\gamma}_k^{(3)}$, so that, for example, the effect of treatment $a = 2$ for a patient with predictive covariate vector $\boldsymbol{z}$ is $d^{(2)}\boldsymbol{z}'\boldsymbol{\gamma}_k^{(2)}$. Assuming the outcomes are conditionally independent between patients, we use the endpoint likelihoods

$$Y_{i1}|\eta_{i1}, \sigma^2 \sim \text{Normal}\left(\eta_{i1}, \sigma^2\right), \quad Y_{i2}|\eta_{i2} \sim \text{Bernoulli}\left(\text{logit}^{-1}\,\eta_{i2}\right), \tag{13}$$

with $\eta_{ik} = \boldsymbol{x}_{ik}'\boldsymbol{\beta}_{k*} + d^{(a_i)}\boldsymbol{z}_{ik}'\boldsymbol{\gamma}_{k*}^{(a_i)}$. We use the prior $\sigma^2 \sim \text{InverseGamma}(0.001, 0.001)$, $\beta_{kp} \sim \text{Normal}\left(0, 10^4\right)$, $\gamma_{k1}^{(a)} \sim \text{Normal}\left(0, 10^4\right)$ for $a > 0$, $\gamma_{kp}^{(a)} \sim \text{Normal}\,(0, 1)$ for $a > 0$ and $p > 1$, $d^{(2)} \sim \text{Uniform}(0, 1)$, and $d^{(1)}|d^{(2)} \sim \text{Uniform}(0, d^{(2)})$. Here we shrink the treatment-covariate interactions to reflect the common

prior belief that such interactions are usually small, and to obtain less variable estimates of conditional treatment effects, but leave the priors for the prognostic effects and baseline treatment effect vague. A sensitivity analysis without such shrinkage did not yield qualitatively different results.

Before using our proposed methods, we analyze the data through a more standard approach. We use a Bayesian model and analysis, though with non-informative priors that correspond to a frequentist analysis. Because our aim is to discuss treatment-covariate interactions, which the study was not powered to detect, we decrease the nominal credible level to 50%. To make the approaches comparable, we will use the same credible level for our proposed analysis. All models are fit with 10,000 Gibbs sampler iterations after 1000 burn-in iterations. We first test the overall effects by removing all γ parameters from the model except the $\gamma_{k1}^{(a)}$, which then correspond to the overall treatment effects versus the placebo. In this analysis, there is a significant overall efficacy difference between the active control and placebo, and between the test treatment and placebo. There are no significant safety differences nor an efficacy difference between the active control and the test treatment.

We continue with a standard subgroup analysis, returning all γ parameters to the model and using minimally informative priors. At the 50% nominal credible level we find significant interactions between the test-placebo efficacy difference and all covariates; and between the test-placebo safety difference and baseline severity and age. We also find significant interactions between the test-active control efficacy difference and baseline severity and carrier status; and between the test-active control safety difference and sex and age. Using a Bonferroni-corrected α-level of $0.50/4$ to account for the four treatment-by-covariate interaction tests per treatment and endpoint (we aren't concerned with multiplicity of endpoints or treatments), we are left with only the interaction of the test-placebo efficacy difference with baseline severity and sex as significant.

We now estimate the average treatment effect of the test treatment versus the placebo in subgroups produced according to the significant interactions (post-Bonferroni) we identified. The treatment effect remains significant in a high-severity (> 22, sample median) subgroup and a low-severity (≤ 22) subgroup, but when grouping by sex, there is a significant effect in males but not females. When the population is divided into four subgroups according to sex $\times$ severity, both male subgroups and neither female subgroup show a significant effect. From this standard subgroup analysis, we get the general idea that the male patients are the primary drivers of the treatment effect versus placebo. However, it is difficult to precisely determine who benefits

from the treatment over the placebo, and especially what treatment effect exists between the test treatment and the active control.

We now compare the high dose test treatment to the placebo and active control simultaneously with respect to the weak and strong admissibility criteria, e.g. $\Delta_{wa}^{a*}(\boldsymbol{z}_i)$ and $\Delta_{sa}^{a*}(\boldsymbol{z}_i)$. In the former case, the benefiting subgroup is the population for which the test treatment is superior to both the placebo and active control with respect to at least one endpoint *or* is inferior to neither the placebo nor the active control with respect to either endpoint. In the latter case, the benefiting subgroup is the one for which the test treatment is superior to both the placebo and active control with respect to at least one endpoint *and* is inferior to neither the placebo nor the active control with respect to any endpoint. The criteria we select for superiority are a difference in expected change in disease severity of greater than $\delta_1 = 0$ and a log odds ratio of adverse event occurrence of less than $-\delta_2 = 0$. The criteria for noninferiority are a difference in expected change in disease severity of greater than $\varepsilon_1 = -0.5$ (standard deviations of the response) and a log odds ratio of adverse event occurrence of less than $-\varepsilon_2 = 0.18$ (corresponding to an odds ratio of approximately 1.20). Signs are switched for δ_2 and ε_2 because we want *reductions* in risk. The model is fit using 100,000 MCMC iterations after 10,000 burn-in iterations. Because of the high memory requirements of constructing credible subgroups over a continuous covariate space, every 10th iteration is used for the computation.

Figure 2 shows individual single-arm single-endpoint credible subgroups plots for each treatment-endpoint combination using $\alpha = 0.50$ for illustration. Because the thresholds for benefit differ from the thresholds for noninferiority, there are in fact two pairs of credible subgroups for each treatment-endpoint combination—one for benefit and one for noninferiority. Letting (D, S) and $(\mathrm{D}', \mathrm{S}')$ be the pairs for benefit and noninferiority, respectively, we have $\mathrm{D} \subseteq \mathrm{D}' \subseteq \mathrm{S} \subseteq \mathrm{S}'$. The upper-left sub-figure shows that males with high disease severity tend to benefit from the test treatment versus the placebo, but in the bottom left sub-figure we detect more non-inferiority in female and low severity patients versus the active control. This hints that the active control and the test treatment may both favor male and high-severity patients relative to the placebo, but that the active control does so to a larger degree, perhaps due to more activity of a similar mechanism. The right-hand side of the figure indicates mostly uncertainty in the relative safety profiles of the treatments, though it appears that female carriers are the most promising for non-inferiority to the active control.

Figure 3 shows credible subgroup pairs for weak and strong admissibility (via the direct methods) against both the placebo and active control. The left

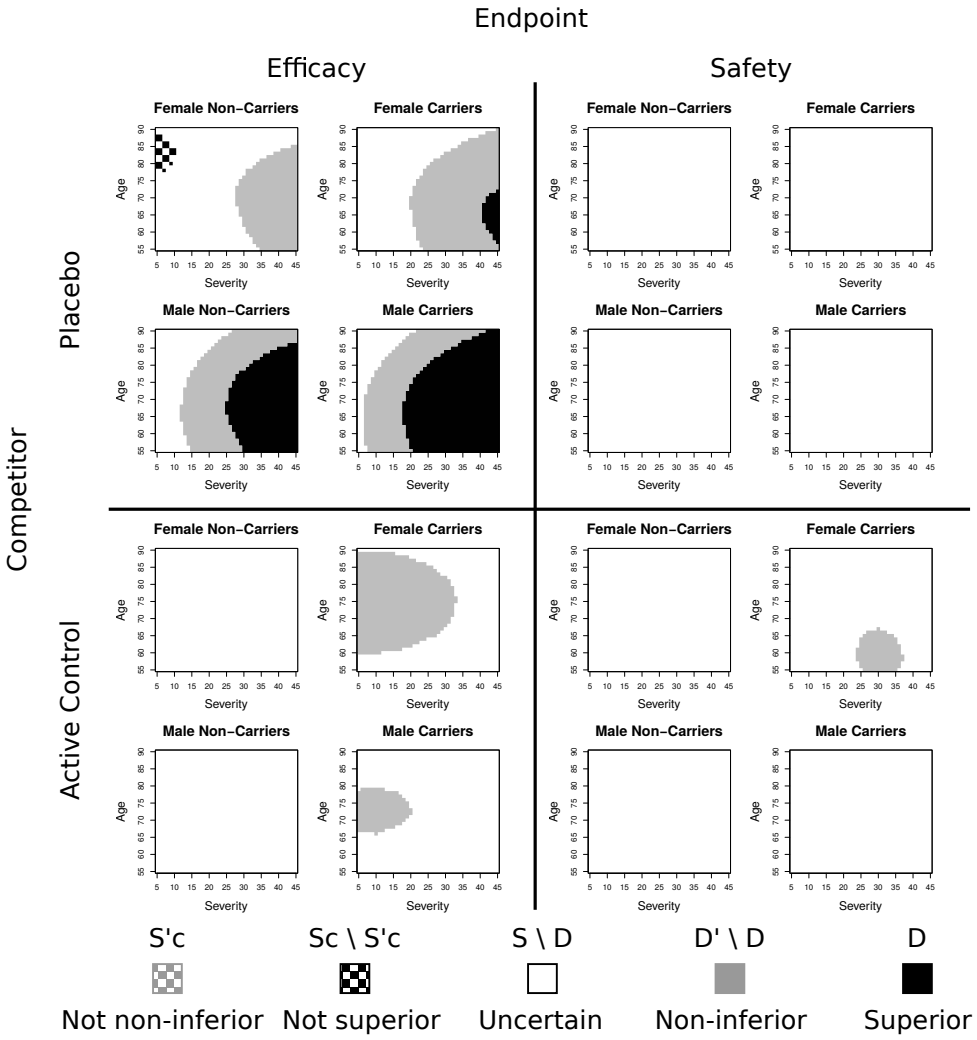

FIG 2. *Individual 50% credible subgroup pairs plotted over the covariate space. Each large quadrant contains a plot of the credible subgroup pairs for that endpoint-competitor combination, with subgroups for noninferiority* $(\mathrm{D}', \mathrm{S}')$ *and superiority* (D, S) *overlaid.*

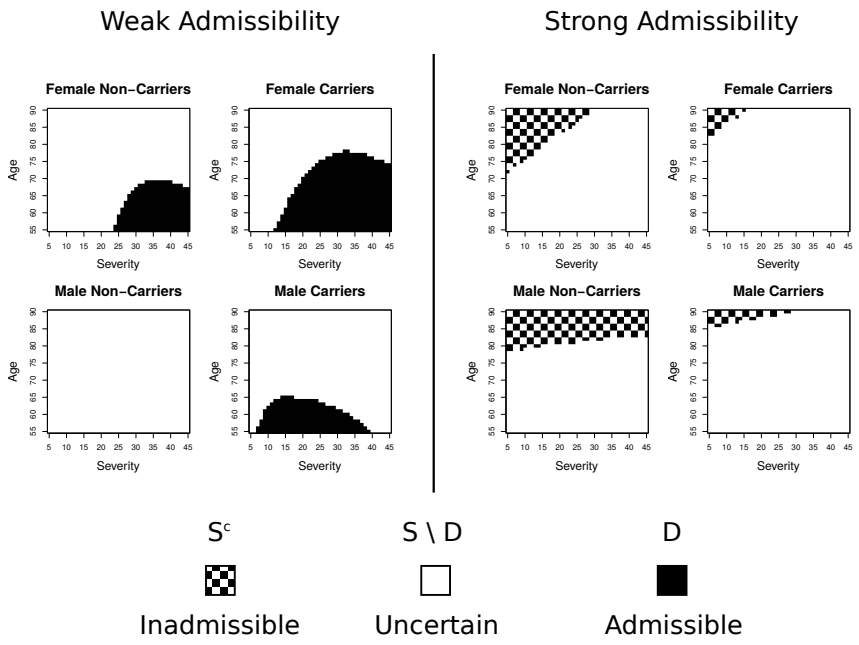

FIG 3. *Admissibility 50% credible subgroup pairs (combined endpoints, versus both placebo and active control) plotted over the covariate space. Left: weak admissibility. Right: strong admissibility.*

sub-figure shows that the exclusive credible subgroup for weak admissibility primarily contains younger patients, and is more present in females and carriers. The features of the weak admissibility credible subgroup plot appear (judging by Figure 2) to come primarily from the efficacy endpoint. The right sub-figure shows that the test treatment is not strongly admissible over an area generally opposite to that over which the test treatment is weakly admissible: older patients, especially males and non-carriers. Though the credible level used is too low to claim conclusive results (e.g., for a regulatory submission), the results provide evidence that the treatment effect is not homogeneous, and indicate which subgroups show promise for appropriately-powered studies in the future.

The model was also fit with spike-and-slab priors similar to George and McCulloch (1993) for variable selection: the $\text{Normal}(0, 1)$ priors for the treatment-covariate interactions were exchanged for $\frac{1}{10}\delta(0)+\frac{9}{10}\text{Normal}(0, 10^4)$ mixture distributions, where $\delta(0)$ is a point mass at 0. The resulting individual credible subgroups by endpoint and arm as in Figure 2 exhibited much more homogeneity: the test treatment was superior to the placebo for all patients with respect to efficacy and indeterminate with respect to safety. Against the active control, the test treatment was noninferior with respect to efficacy for patients with baseline severity < 31, and for women younger than 80 and men younger than 70 with respect to safety. The test treatment was weakly admissible for all patients, and strong admissibility was indeterminate for most patients, and negative for the patients with the highest baseline severity (near 45).

4. Discussion. The medical community recognizes the need to consider the characteristics of individual patients when deciding avenues of treatment. In addition to baseline covariates that are predictive of treatment effects with respect to single endpoints, it is also necessary to consider differences in individuals' preferences that may lead different patients to differentially value endpoints. For example, one patient may pursue the most efficacious treatment while another prefers a treatment with side effects that minimally affect quality of life.

The concept of admissibility provides a utility function–free approach to summarizing treatment effects with respect to multiple endpoints, and admits a natural extension to trials with more than two arms. In this paper we have also examined multiple definitions of admissibility in the clinical trial context, as well as estimators which do and do not adjust for the multiplicity of endpoints so that Type I error may be controlled. Finally, the credible subgroups method of Schnell et al. (2016) provides a natural im-

plementation for admissibility ideas by also adjusting for the multiplicity of covariate points, and we generalize the previously published method to handle settings outside of the normal linear model by requiring only a sample from the joint posterior of personalized treatment effects, allowing the consideration of generalized linear and other more sophisticated models.

While the confidence levels used in Figures 2 and 3 are too low for our results to be considered definitive, it is important to note that they are based on data from a study not powered to deliver simultaneous inference on multiple endpoints across arbitrary subgroups defined by up to four different covariates. So while these results are far from convincing for final regulatory approval, they do provide valuable information about the sort of enrollees that should be sought for future, more focused subgroup-confirmatory trials. For instance, the weak admissibility portion of Figure 3 suggests younger females with more severe dementia would make good candidates, whereas the strong admissibility portion discourages enrollment of older patients, particularly those with less severe dementia. Used in this way, our methods essentially become a useful tool for enrichment designs (Peace and Chen, 2010).

Finally, the relationship between identifying admissible treatments in the development and regulatory context treated here, and the single-patient focus of the dynamic treatment regime context, present an interesting duality between decisions made in relation to a given treatment versus a given patient. For example, developer-sponsored clinical trials may aim to secure regulatory approval for therapies in specific subpopulations, and optimal treatment regimes may subsequently be constructed on a per-patient basis from available treatments using the concepts of admissibility, which are similar to the non-domination criteria used in Laber, Lizotte and Ferguson (2014). Attempts toward unifying development, regulatory, and patient-care contexts may represent a promising avenue for future research.

Acknowledgments. We thank the Editor, Assistant Editor, and reviewers for their helpful comments. This work was supported by AbbVie, Inc; and the National Cancer Institute [1-R01-CA157458-01A1 to PMS and BPC]. AbbVie contributed to the design, research, interpretation of data, reviewing, and approving of this publication.

SUPPLEMENTARY MATERIAL

Supplement to "Subgroup Inference for Multiple Treatments and Multiple Endpoints in an Alzheimer's Disease Treatment Trial" (doi: COMPLETED BY THE TYPESETTER; .zip). The online supplement

contains proofs related to the construction of simultaneous credible bands (as in (4) and (7)), as well as the data and scripts required to reproduce the simulations and analyses presented above.

References.

Almirall, D., Lizotte, D. J. and Murphy, S. A. (2012). Comment. *Journal of the American Statistical Association* **107** 509–512.

Berger, J. O., Wang, X. and Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* **24** 110–129.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL. CRC Press.

Burns, A. and Iliffe, S. (2009). Alzheimer's disease. *British Medical Journal* **338**.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association* **93** 935–948.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4** 266–298.

Dixon, D. O. and Simon, R. (1991). Bayesian subset analysis. *Biometrics* 871–881.

Foster, J. C., Taylor, J. M. and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30** 2867–2880.

Freidlin, B., Jiang, W. and Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research* **16** 691–698.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11** 7872–7878.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.

Laber, E. B., Lizotte, D. J. and Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70** 53–61.

Lizotte, D. J., Bowling, M. and Murphy, S. A. (2012). Linear fitted-Q iteration with multiple reward functions. *Journal of Machine Learning Research* **13** 3253–3295.

Lizotte, D. J. and Laber, E. B. (2016). Multi-objective Markov decision processes for data-driven decision support. *Journal of Machine Learning Research* **17** 1–28.

Peace, K. E. and Chen, D.-G. D. (2010). *Clinical Trial Methodology*. Boca Raton, FL: Chapman and Hall/CRC Press.

Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* **21** 2917–2930.

Schnell, P. M., Tang, Q., Offen, W. W. and Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72** 1026–1036.

Schnell, P. M., Tang, Q., Müller, P. and Carlin, B. P. (2017). Supplement to "Subgroup Inference for Multiple Treatments and Multiple Endpoints in an Alzheimer's Disease Treatment Trial".

Sivaganesan, S., Laud, P. W. and Müller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Pólya Urn scheme. *Statistics in Medicine* **30** 312–323.

NIMBLE Development Team (2015). NIMBLE: An R Package for Programming with BUGS models, Version 0.4.

Thall, P., Sung, H. and Estey, E. (2002). Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials. *Journal of the American Statistical Association* **97** 29–39.

Thall, P. F., Logothetis, C., Pagliaro, L. C., Wen, S., Brown, M. A., Williams, D. and Millikan, R. E. (2007). Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. *Journal of the National Cancer Institute* **99** 1613–1622.

Uusipaikka, E. (1983). Exact confidence bands for linear regression over intervals. *Journal of the American Statistical Association* **78** 638–644.

Xu, Y., Trippa, L., Müller, P. and Ji, Y. (2016). Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences* **8** 159–180.