Received: 25 September 2023 | Revised: 15 February 2024 | Accepted: 2 April 2024

DOI: 10.1002/sim.10083

RESEARCH ARTICLE

Statistics in Medicine WILEY

Finding the best subgroup with differential treatment effect with multiple outcomes

Beibo Zhao[1] | **Jason Fine**[2] | **Anastasia Ivanova**[1]

[1]Department of Biostatistics, CB #7420, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

Correspondence
Anastasia Ivanova, Department of Biostatistics, CB #7420, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420, USA.
Email: aivanova@bios.unc.edu

Funding information
National Heart, Lung, and Blood Institute, National Institutes of Health, Grant/Award Number: U24 HL138998

Precision medicine aims to identify specific patient subgroups that may benefit the most from a particular treatment than the whole population. Existing definitions for the best subgroup in subgroup analysis are based on a single outcome and do not consider multiple outcomes; specifically, outcomes of different types. In this article, we introduce a definition for the best subgroup under a multiple-outcome setting with continuous, binary, and censored time-to-event outcomes. Our definition provides a trade-off between the subgroup size and the conditional average treatment effects (CATE) in the subgroup with respect to each of the outcomes while taking the relative contribution of the outcomes into account. We conduct simulations to illustrate the proposed definition. By examining the outcomes of urinary tract infection and renal scarring in the RIVUR clinical trial, we identify a subgroup of children that would benefit the most from long-term antimicrobial prophylaxis.

KEYWORDS
cross-validation, multiple outcomes, predictive biomarker, subgroup analysis, treatment effect heterogeneity

1 | INTRODUCTION

With the increasing popularity of personalized medicine, there has been a growing interest in clinical communities to do more than just establish the overall efficacy of a treatment at the end of a randomized clinical trial through demonstrating a favorable treatment effect in all patients. There is now a focus on exploring the heterogeneity of treatment effects. This involves discovery of patient subgroups that may have more pronounced responses than the overall population (perhaps even in absence of an overall effect), and identification of baseline biomarkers (ie, genomic, clinical, socio-demographic, and other covariates) with strong predictive properties. The goal of this practice is to maximize the overall treatment benefit by treating patients who are likely to respond to the intervention and to minimize unnecessary exposure to potentially harmful side-effects by not treating patients who are unlikely to respond. A major application of conducting post-hoc subgroup discovery with principled data-driven methodologies is to provide guidance for optimizing design, sample size allocations, and the power in future trials where confirmatory subgroup analyses can be performed.[1]

There are two phases in a subgroup analysis: identification and confirmation. The typical goal of the identification phase is to estimate the best subgroup as a function on a set of predictive baseline biomarkers. Within the frequentist framework and under the single-outcome setting, the definition of the best subgroup typically involves the conditional average treatment effect (CATE), that is, the average treatment effect conditioned on biomarkers. The subgroup is often constructed with binary or dichotomized continuous/ordinal biomarkers, with the process of dichotomization based on pre-specified or estimated cutoff values.

Some definitions focus solely on CATE. Imai and Ratkovic select the subgroup comprising all individuals with a CATE >0.[2] Huber et al.[3] and Foster et al.[4] extend the definition to all individuals with a CATE greater than the average treatment effect in the population, or a predefined minimum clinical benefit threshold. Tian et al.[5] simply dichotomize the population and designate the half comprising all individuals with a CATE higher than an estimated threshold as the best subgroup.

Some definitions consider the size of the subgroup. Lai et al.[6,7] define the best subgroup as the cutoff of a single biomarker through maximizing a utility function that represents the Kullback-Leibler information number, which under the assumption of equal variances of treatment effects in all subgroups, is equivalent to maximizing the power of testing the CATE for all individuals within that subgroup (ie, the CATE in the subgroup), and same as maximizing the product of subgroup prevalence raised to the power of 0.5 and the CATE in the subgroup. Joshi et al. adopt this definition and consider subgroup prevalence raised to the power of 0.75 for the design purposes and 0.5 for estimation.[8,9]

Some definitions include the sample size. Chen et al.[10] define the best subgroup as the one comprising all individuals with greater CATE and stronger statistical significance of testing the CATE than its complement. Huang et al.[11] adds to Chen et al. by including a pre-specified threshold to the *P*-value of testing the CATE in the subgroup and describing the best subgroup as intersections of half-lines for some subset of predictive biomarkers. Zhang et al.[12] propose a utility (which they refer to as expected gain) as the product between a specified function of subgroup prevalence and the power of testing the CATE in the subgroup. When the specified function does not exist or is a constant, their objective is simply power. These definitions focus on balancing the trade-off between the size and the CATE in the subgroup. The enrichment design from Wang et al.[13] also utilizes this trade-off.

Many methods do not clearly define the best subgroup through true population parameters and power but through estimated quantities from their proposed estimation methods. Su et al.[14,15] follow the Classification and Regression Trees (CART) approach and define the best subgroup as cutoffs of biomarkers through optimizing an objective function that is equivalent to minimizing the *P*-value from testing the biomarker-treatment interaction in their specified model. Seibold et al.[16,17] define the best subgroup as the partition of the biomarker space through minimizing the sum of the negative log-likelihoods if their fitted parametric model. Xu et al.[18] use the sign of their estimated linear predictors to define the best subgroup through minimizing their proposed objective function with two imposed penalty terms. Chen et al.[19] defines the best subgroup through finding a score function as a linear combination of biomarkers that minimizes their proposed quantity.

In clinical trials, multiple outcomes are sometimes considered to evaluate the efficacy of a treatment. This work is motivated by two clinical trials. The Precision Interventions for Severe and/or Exacerbation-prone Asthma (PrecISE) study (ClinicalTrials.gov Identifier: NCT04129931). The PrecISE study is a clinical trial to investigate five novel therapies for patients with severe asthma.[20] In the PrecISE study the effect of a treatment is being evaluated based on the following three endpoints: forced expiratory volume in one second percent predicted, the 6-item Juniper Asthma Control Questionnaire score, and the CompEx events, a composite endpoint that includes asthma exacerbations.[21] The goal of the study is to investigate the efficacy of the each of the five interventions compared to placebo with respect to the three endpoints. Another objective is to estimate the best subgroup for each intervention as a function of a set of baseline biomarkers. Including multiple outcomes, potentially of different types, can maximize the information to answer the question of efficacy and likely lead to better subgroup estimation after the trial, than using a single outcome. The second trial is the Randomized Intervention for children with Vesicoureteral Reflux (RIVUR) trial. This was a randomized, placebo-controlled trial aimed at evaluating the efficacy of long-term antimicrobial prophylaxis in preventing recurrences of urinary tract infection (UTI) and renal scarring in children with vesicoureteral reflux (VUR).[22] Routine use of long-term antimicrobial prophylaxis is controversial because it may lead to the development of antibiotic resistance and alterations of microbiome. It is valuable to identify smaller subgroups[23] of children who would benefit from antimicrobial prophylaxis using not just urinary tract infection but renal scarring as well.

Not many prior publications address the problem of defining the best subgroup in a multiple-outcome setting. Loh et al.[24,25] extend their previously published tree-based classification and regression method, Generalized, Unbiased Interaction Detection and Estimation (GUIDE), to multiple outcomes. In the case of a single outcome, GUIDE first tests each biomarker for interaction with treatment and chooses the most significant biomarker to split on by maximizing a 1-df chi-squared value transformed from a F-statistic for the "pure error" lack-of-fit test. The best subgroup is the one that minimizes the sum of squares of residuals in the child node. In the case of multiple outcomes, GUIDE is applied to each outcome at a time and the best subgroup is the one that maximizes the sum of chi-squared values over all outcomes. This method can be applied to continuous and binary outcomes but not censored time-to-event outcomes. Zhang et al.[26] extend their previous non-parametric method that searches for subgroup membership scores by maximizing a value function

which reflects the biomarker-treatment interaction for a single time-to-event outcome, to multiple outcomes based on win difference. However, this method is under the framework of individualized treatment rules (identify best treatment for a patient) rather than that of subgroup identification (identify best patients for a treatment), with no variable selection.

In this article, in the case of multiple outcomes, we propose a definition of the best subgroup that offers a trade-off between the subgroup size and the CATE in the subgroup, with respect to each of the outcomes. The definition includes optimal weights for the outcomes to achieve the maximum utility or, equivalently, the best power. We adapt several established subgroup identification methods to estimate the best subgroup using the proposed definition. We use simulations to evaluate the feasibility of using the proposed definition and apply it to the RIVUR trial.

The rest of the article is organized as follows. Section 2 gives the proposed definition of the best subgroup under a multiple-outcome setting. Section 3 describes the subgroup identification methods to estimate the best subgroup, the performance metrics to evaluate the estimation accuracy, and the testing procedure of the CATE of the estimated best subgroup. Section 4 describes the simulation setting and presents simulation results. Section 5 illustrates the real-data application to the RIVUR trial. Section 6 presents the conclusion.

2 | METHODOLOGY

2.1 | Setup

Consider a post hoc analysis of a parallel group clinical trial where n patients are randomized to either the treatment arm or the control arm with probability 0.5. Let T denote the treatment indicator, 0 for control, 1 for treatment; Let $\boldsymbol{X} = (X_1, \dots, X_M)$ denote M continuous or discrete biomarkers of interest measured at baseline prior to treatment. Let $\boldsymbol{Y} = (Y_1, \dots, Y_J)$ denote J continuous, binary, or time-to-event outcomes. We assume that the observed data consist of n independent and identically distributed copies drawn from the joint distribution of random variables $(\boldsymbol{Y}, \boldsymbol{X}, T)$. For the i-th patient, we denote their m-th observed baseline biomarker as $x_{m,i}$, $m = 1, \dots, M$, j-th observed outcome as $y_{j,i}$, $j = 1, \dots, J$, and treatment received as T_i, $i = 1, \dots, n$. Hence, $\boldsymbol{x}_i = \left(x_{1,i}, \dots, x_{M,i}\right)$ is the M-dimensional vector of observed baseline biomarkers, and $\boldsymbol{y}_i = \left(y_{1,i}, \dots, y_{J,i}\right)$ is the J-dimensional vector of observed outcomes. Let $\boldsymbol{H}(\boldsymbol{X}, \boldsymbol{T}) : \boldsymbol{\mathcal{R}}^{M+1} \to \boldsymbol{\mathcal{R}}^{p}$ be a p-dimensional function of baseline biomarkers and treatment indicator, including the intercept.

Consider a single outcome $Y = Y^{(T)}$, with $Y^{(1)}$and $Y^{(0)}$ be the potential outcome if the patient receives treatment 1 and 0, respectively. When the outcome is time-to-event, we assume that the outcome Y is a pair of random variables $(Q, \delta) = \{\tilde{Q} \wedge C, I(\tilde{Q} < C)\}$, where $\tilde{Q}$ is the event time, C is the right-censoring time, $\tilde{Q} \perp C$, and δ is the right-censoring indicator. Without loss of generality, assume higher Y values for continuous and binary outcomes, and higher $\tilde{Q}$ values (longer latent event time) for time-to-event outcome, indicate more favorable clinical results.

The fundamental problem of inferring causal relationships between biomarkers and outcomes in a parallel group trial setting is that we cannot simultaneously observe $Y^{(1)}$ and $Y^{(0)}$. Under this framework, let subgroup $S \subseteq \mathbb{X}$ where $\mathbb{X}$ is the support space spanned by $\boldsymbol{X}$. S is defined by a rule, which selects a subset of the overall population based on $\boldsymbol{X}$. For example, $S = I\{X_1 > c\}$. Let $\pi(S)$ denote the prevalence of S; $\Delta(S)$ denote the CATE of S. For the continuous outcome, we define $\Delta(S)$ as:

$$\Delta(S) = E\left(Y^{(1)} \mid S\right) - E\left(Y^{(0)} \mid S\right).$$

For the binary outcome, we define $\Delta(S)$ as:

$$\Delta(S) = \log\left(\frac{1 - \left(p_S^{(1)} - p_S^{(0)}\right)}{1 + \left(p_S^{(1)} - p_S^{(0)}\right)}\right),$$

where $p_S^{(1)} = E\left(Y^{(1)} \mid S\right); p_S^{(0)} = E\left(Y^{(0)} \mid S\right)$.

For the time-to-event outcome, assuming balanced censoring rates between the two arms and proportional hazards, we define $\Delta(S)$ following that from Tian et al.[5] as:

$$\Delta(S) = -\frac{1}{2}\log\left\{\frac{E\left[\Lambda_0\left(\tilde{Q}^{(1)}\right)|\, S\right]}{E\left[\Lambda_0\left(\tilde{Q}^{(0)}\right)|\, S\right]}\right\},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is a monotone increasing function; $\lambda(t|\cdot)$ is the hazard function for survival time $\tilde{Q}$.

Following the work of Tian et al., the presented expressions of the CATE describe the biomarker-specific treatment effect, which captures the heterogeneity in treatment effects across the population. These are derived from a straightforward working model that yields the modified covariate estimator, providing reliable estimates even when the assumptions of the working model may not hold. An alternative approach to construct CATE with a binary outcome is through the relative risk, as provided by Tian et al. in their Supplementary materials. We do not consider this approach due to the possible complexity of having zero in the denominator.

2.2 | Definition of the best subgroup in multiple-outcome setting

For a single outcome, define the best subgroup, S_{true}, as the subgroup that maximizes utility $U(S) = \Delta(S)\pi(S)^{0.5}$ over all possible subgroups in the subgroup space, that is, $U(S_{true}) = \max_S U(S)$. The utility $U(S)$ allows for a trade-off between the subgroup size and the CATE of subgroup.[6,7] This definition is also beneficial in trials with adaptive enrichment since this utility is proportional to the test statistics to test for the CATE assuming equal variances of treatment effects in all subgroups. Maximizing this utility is equivalent to maximizing the power to test for the CATE due to the exact correspondence between power and test statistics curves for subgroups defined by the minimum CATE threshold.[6,7,27] We extend this definition to the case of multiple outcomes. For the multiple-outcome setting, we define the best subgroup S_{true} and the optimal weight vector $\mathbf{w}_{true}$ as the pair that jointly maximizes the utility $U(S, \mathbf{w}) = w_1U_1(S) + \ldots + w_JU_J(S)$ over all possible subgroups in the subgroup space and all possible weight vectors $\mathbf{w} = (w_1, \ldots, w_J)$ such that $\sum_{j=1}^J w_j^2 = 1$. $U_j(S) = \Delta_j(S)\pi(S)^{0.5}$ is the utility of the subgroup with respect to an outcome $j = 1, \ldots, J$. The weight vector assigns varying levels of emphasis to the outcomes. If there is no unique best subgroup by this definition, we say that there is no subgroup with differential treatment effect in the population. This definition allows for a trade-off between the subgroup size and the CATE in the subgroup with respect to each outcome under the multiple-outcome setting. Typically, the optimal weights are not known in advance. The weights are estimated from data through searching across a large set of possible weight vectors, at the same time as the best subgroup is being estimated.

3 | SUBGROUP ESTIMATION

3.1 | Subgroup identification methods

To estimate S_{true}, a pool of possibly overlapping candidate subgroups, which includes the full population, is first generated. The estimated best subgroup, $\widehat{S}$, is the subgroup that maximizes the proposed form of utility over all candidate subgroups in the pool. We adapt several existing subgroup identification methods from the single-outcome setting to the multiple-outcome setting to generate this pool.

Penalized regression is a global outcome modelling method that explicitly incorporates model selection and provides a comprehensive characterization of treatment effect heterogeneity through estimating the expected outcome function. Functions of interest can be defined as linear expansions of basic functions.[28] For example, baseline biomarkers as the main effects to represent the prognostic effects, their interactions with treatment to represent the predictive effects, and the treatment indicator. Proper constraints (ie, penalty functions) are used to perform parameter estimation simultaneously with automatic variable selection. When lacking knowledge of how potentially predictive biomarkers interact with each other to influence the outcome, the complexity control from the penalty terms supports complex fits that expand the biomarker space.[1] We generate the pool of candidate subgroups through thresholding linear predictors from fitting a penalized regression with a lasso-type regularization, the overlapping-group exponential lasso (OG-EL) penalty. The resulting candidate subgroups are defined as inequalities between a linear function of biomarkers and a cut-off value. There are no known theoretical results on the asymptotic oracle property or the finite-sample oracle inequality of OG-EL penalty. The details of this OG-EL method are in Supplement.

We modify CART to use a splitting rule that is based on the proposed best subgroup definition, that is, maximizing the form of utility at each split.[29] For the censored time-to-event outcome, we adapt the survival tree algorithm with the splitting rule developed by Leblanc and Crowley[30] who use the first step of a full likelihood estimation under proportional hazards assumption. A complexity parameter of 0.01, and a minimal tree size of 10% of the sample size is implemented. All pruned splits form the pool of candidate subgroups. The resulting candidate subgroups are defined as combinations of biomarker cut-offs.

We implement the regression-based treatment-biomarker interaction modelling method with modified biomarkers (MOD) from Tian et al.,[5] which is applicable to all three outcome types we consider. The linear combination of estimated treatment-biomarker interaction coefficients and biomarkers is used to calculate individualized treatment effects for each outcome. The observations are then dichotomized into high and low treatment-effect groups according to the median value, with the high group from each outcome forming the pool of candidates defined as inequalities between a linear function of biomarkers and a cut-off value.

3.2 | Testing under the multiple-outcome setting

The estimated best subgroup $\widehat{S}$ is obtained from a pool of candidate subgroups. We are naturally interested in testing the joint null hypothesis that there is no CATE in the subgroup for any of the outcomes. Conversely, the joint alternative hypothesis posits that there exists a non-zero CATE in the subgroup for at least one of the outcomes. It is a challenging task to obtain a proper *P*-value to test for the CATE with respect to multiple outcomes. In the case of a known vector of weights and a small number of binary biomarkers, an ordinal biomarker or a single biomarker with several possible cutoffs, the adjusted *P*-value can be obtained from the joint distribution of the test statistics for possible subgroups.[31,32] Similarly, for a given subgroup, if only several possible sets of weights are being considered, the adjusted *P*-value can be obtained using the Bonferroni approach or other methods to adjust for multiplicity. When many possible candidate subgroups and weights are being considered, and no independent test datasets are available, resampling-based approaches may be the only feasible approach within the frequentist framework, and they often involve replicating the entire subgroup identification strategy, including estimation of any data-driven tuning parameter, in each resampled dataset.[1,4,12,27,33-36]

A re-substitution approach is to simply estimate $\widehat{S}$ from the whole sample and proceed to inference. This practice often leads to the presence of re-substitution bias, resulting in over-estimation of treatment effects and an inflated type I error rate.[33,37] Here we use a *K*-fold cross-validation procedure similar to the cross-validated version in Freidlin et al.[33,38] that we adapt to use with multiple outcomes.

First, split the sample into $k = 1, ..., K$ cohorts of equal size. Second, to patients not in cohort k, apply the subgroup identification method and estimate $\left(\widehat{S}_{\{-k\}}, \widehat{\mathbf{w}}_{\{-k\}}\right)$ which jointly maximize the proposed utility for the J outcomes, where $\widehat{\mathbf{w}}_{\{-k\}} = \left(\widehat{w}_{\{-k\},1}, \ldots, \widehat{w}_{\{-k\},J}\right)$. Third, apply $\left(\widehat{S}_{\{-k\}}, \widehat{\mathbf{w}}_{\{-k\}}\right)$ to patients in cohort k. To adapt to testing with multiple outcomes, we apply the weighted *Z*-method in each cohort so information from nulls of individual outcomes are combined to determine if the joint null should be rejected.[39] Let $Z_{k,j,\widehat{S}_{\{-k\}}}$ denote the test statistic for testing $\Delta_j\left(\widehat{S}_{\{-k\}}\right)$ of outcome j in $\widehat{S}_{\{-k\}}$ in cohort k. We compute the weighted Z for the J outcomes in cohort k, $Z_{k,\widehat{S}_{\{-k\}}} = \frac{\sum_{j=1}^{J} \widehat{w}_{\{-k\},j} Z_{k,j,\widehat{S}_{\{-k\}}}}{\sqrt{\sum_{j=1}^{J} \widehat{w}^2_{\{-k\},j}}}$. Repeat the process K times over all K cohorts.

Consider the following naïve test statistic that combines statistics from the K cohorts. $\tilde{Z}_{\widehat{S}} = \sqrt{1/K} \sum_{k=1}^{K} Z_{k,\widehat{S}_{\{-k\}}}$. This test statistic does not preserve the type I error rate because statistics are correlated. A permutation-based test $Z_{\widehat{S}}$ is derived to provide a valid *P*-value.[33] Permuted data sets are constructed by randomly permuting treatment assignments, then the *K*-fold cross-validation procedure is repeated for each permuted data set, yielding test statistic $\widehat{Z}_{\widehat{S}}$. The permutation *P*-value can be computed as:

$$\frac{1 + \#\text{ of permutations with } \widehat{Z}_{\widehat{S}} \geq \tilde{Z}_{\widehat{S}}}{1 + \#\text{ of permutations}}.$$

In Figure 1, we visualize the subgroup estimation and testing procedure with a flowchart.

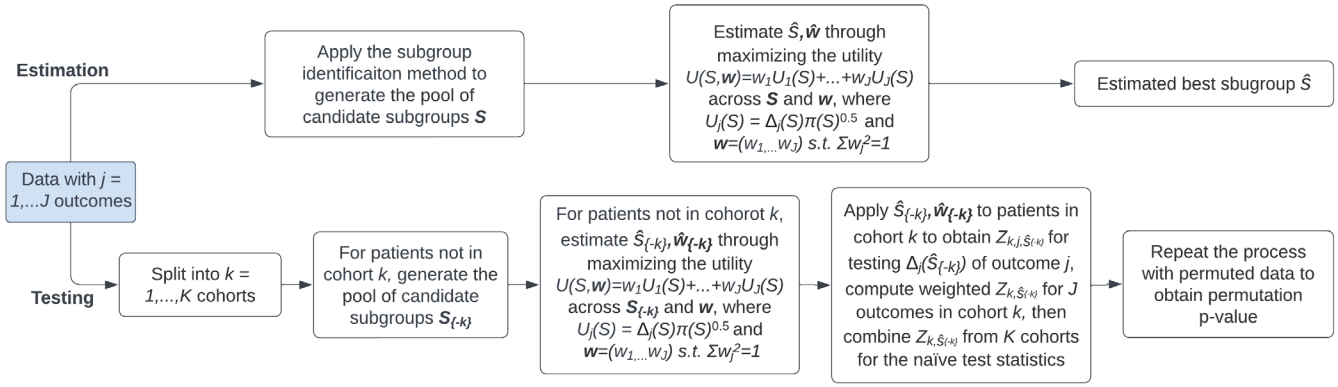

FIGURE 1 Flowchart of the subgroup estimation and testing procedure.

3.3 | Performance Metrics

We assess the performance of the proposed best subgroup definition in subgroup identification methods through subgroup estimation accuracy, variable selection accuracy, and empirical power. To evaluate subgroup estimation accuracy, we modify the metric introduced in Joshi et al.[9] for the multiple-outcome setting. We propose a metric $\%U$ defined as the percentage of the utility ratio $U(\widehat{S}, \widehat{\mathbf{w}})/U(S_{true}, \mathbf{w}_{true})$ in which $U(\widehat{S}, \widehat{\mathbf{w}})$ is the true utility with an estimated optimal weight vector $\widehat{\mathbf{w}}$ and an estimated best subgroup $\widehat{S}$; $U(S_{true}, \mathbf{w}_{true})$ is the true utility of the best subgroup S_{true} with the optimal weight vector $\mathbf{w}_{true}$. The higher the $\%U$, the more accurate the estimation. If $\widehat{S}$ matches completely with S_{true}, then $\%U$ is 100%. If there is no attempt to find a subgroup, then treatment is recommended for the full population. We denote the corresponding $\%U$ for the full population as $\%U_{all}$. We compute $\%U$ by applying $(\widehat{S}, \widehat{\mathbf{w}})$ and $(S_{true}, \mathbf{w}_{true})$ to a simulated external validation sample of size 10000 in each arm. In addition, we can compare the subgroup prevalence of the estimated subgroup $\pi(\widehat{S})$ and the subgroup prevalence of the true subgroup $\pi(S_{true})$. To evaluate variable selection accuracy, we report the percent of simulation runs where the exact set of biomarkers determining S_{true} is selected in $\widehat{S}$.

4 | SIMULATIONS

We performed Monte Carlo simulations to assess the finite-sample performance of the proposed best subgroup definition in a selection of subgroup identification methods under multiple-outcome settings (with continuous, binary, and censored time-to-event outcomes).

4.1 | Scenarios

We considered $M = 4$ independent $U(0, 1)$ biomarkers and $J = 3$ outcomes, with expected responses under treatment and control arms, $Y_j^{(1)}$ and $Y_j^{(0)}$, defined separately for each outcome j, $j = 1, 2, 3$, in two multiple-outcome settings. We considered a continuous outcome $Y_1^{(T)} \sim N\left(\mu_1^{(T)}, \sigma^2\right)$, a binary outcome $Y_2^{(T)} \sim \text{Binomial}\left(\text{logit}^{-1}\left(u_2^{(T)}\right)\right)$ and a censored time-to-event outcome $Y_3^{(T)} \equiv \left(Q^{(T)}, \delta^{(T)}\right) = \left\{\tilde{Q}^{(T)} \wedge C^{(T)}, I\left(\tilde{Q}^{(T)} < C^{(T)}\right)\right\}$, where $\tilde{Q}^{(T)}$ is the Weibull time-to-event time generated as:

$$\tilde{Q}^{(T)} = \left(-\frac{\log(U[0,1])}{\kappa_q^* e^{\mu_3^{(T)}}}\right)^{\frac{1}{v}},$$

and $C^{(T)}$ is the right-censoring time randomly drawn from an exponential distribution with the mean of $1/\kappa_c^{(T)}$, $\tilde{Q}^{(T)} \perp C^{(T)}$, and $\delta^{(T)}$ is the censoring indicator. We used $\kappa_q = 0.001$, $v = 1$ in latent time generation, and $\kappa_c^{(1)} = 0.005$, $\kappa_c^{(0)} = 0.003$ in censoring time generation to induce ~25% right-censoring rate balanced in both arms.

The mean $\mu_j^{(T)}$ for outcome j with the treatment indicator T was generated from the following change-point model:

$$\mu_j^{(T)} = \alpha_{0,j} + \alpha_{1,j}T + \alpha_{2,j}TI(S_{true}), \tag{1}$$

where $\alpha_{0,j}$ is the common effect; $\alpha_{1,j}$ is the treatment effect independent of S_{true}; $\alpha_{2,j}$ is the differential treatment effect specific to S_{true}.

In the first setting, we generated multiple outcomes independently. In the second setting, we introduced correlations to outcomes of different types through the use of a copula, which allows specification of a multivariate distribution for which the marginal distributions of outcomes are the same as in previous settings. This approach allows modelling the dependence structure between outcomes with known marginal distributions flexibly even if they are of different types.[40] We used a multivariate Gaussian copula so that the dependencies between each pair of outcomes followed a Gaussian distribution, with a correlation of 0.8 between each pair of the outcomes on the same patient given the same intervention, a correlation of 0 between each pair of the outcomes on the same patient given the different intervention.

In each multiple-outcome setting, our models have the following S_{true} for each of the outcomes:

Model 0 (Null): $S_{true} = \emptyset$,
Model 1: $S_{true} = \{\boldsymbol{X} : X_1 > 0.5\}$,
Model 2: $S_{true} = \{\boldsymbol{X} : X_3 + X_4 > 1\}$,
Model 3: $S_{true} = \{\boldsymbol{X} : X_2 > 1/3,\ X_3 + X_4 > \sqrt{0.5}\}$,
Model 4: $S_{true} = \{\boldsymbol{X} : X_1 > 1/5,\ X_2 > 1/6,\ X_3 + X_4 > \sqrt{0.5}\}$,
Model 5 (Full): $S_{true} = \mathbb{U}$,

where $\emptyset$ denotes the empty set, representing the null case that the treatment should not be recommended to anyone; $\mathbb{U}$ denotes the universal set of all patients regardless of their biomarker values, representing the full case that the treatment should be recommended to everyone.

Without loss of generality, for an outcome j, we set $\alpha_{1,j} = 0$, and set $\alpha_{0,j}$, $\alpha_{2,j}$ values varying by outcome types. In Model 1-4, S_{true} have a subgroup prevalence of 0.5, and we chose the parameter values in Equation (1) such that $\Delta(S_{true}) = 0.6$ and 0 outside S_{true}, yielding $U(S_{true}) = 0.6 * \sqrt{0.5} \approx 0.424$. For a continuous outcome, we set $\alpha_{0,j} = 0.1$, $\alpha_{2,j} = 0.6$. For a binary outcome, we set $\alpha_{0,j} = 0.025$, $\alpha_{2,j} = 1.35$ to have 50% success probability overall in the control arm, 65% success probability overall in the treatment arm, and 80% success probability in S_{true} in the treatment arm. For a censored time-to-event outcome, we set $\alpha_{0,j} = 0.05$, $\alpha_{2,j} = 1.1$ to have a median observed event time of ~180 days in the treatment arm, and ~300 days in the control arm, with ~25% right-censored rate in both arms. For each outcome, the power for testing the individual null hypothesis that there is no CATE against the alternative hypothesis that there is a CATE is 85% if the treatment is tested in the overall population and 95% if the treatment is tested in S_{true}. In Model 0, we choose the parameter values in Equation (1) such that $\Delta(S_{true}) = 0$. In Model 5, we choose the parameter values in Equation (1) such that $\Delta(S_{true}) = 0.3$, yielding $U(S_{true}) = 0.3 * \sqrt{1} = 0.3$, leading to an overall power of 85%. In both cases, there is no subgroup with differential treatment effect in the population.

In each multiple-outcome setting, we considered the "all outcomes > 0" scenarios where all three outcomes have the same non-zero effect size $\Delta(S_{true})$, and the optimal weight vector assigns equal weights of $\sqrt{1/3}$ to all outcomes. In this case, the most powerful approach is to test the CATE in the subgroup based on the test statistic equal to the weighted sum of outcome-specific tests. For the "one outcome > 0" scenarios, only one of the three outcomes has the non-zero effect size $\Delta(S_{true})$. The optimal weight vector assigns a weight of 1 to this outcome and 0 to other two outcomes with zero effect sizes. The "one outcome" scenarios serve to assess the influence of the remaining outcomes, which in this context are "noise" outcomes.

4.2 | Simulation results

Simulations were performed in R[41] with 1000 trials with a sample size of 400 (200 in the treatment arm; 200 in the control arm) under each multiple-outcome setting. For all methods, we considered candidate subgroups that were at least 10% of the sample. The optimal weight vector was estimated from data because the relative effect sizes among the outcomes were not known. When estimating power, we increased the number of trials to 5000 for Model 1-5 and to 10000 to estimate

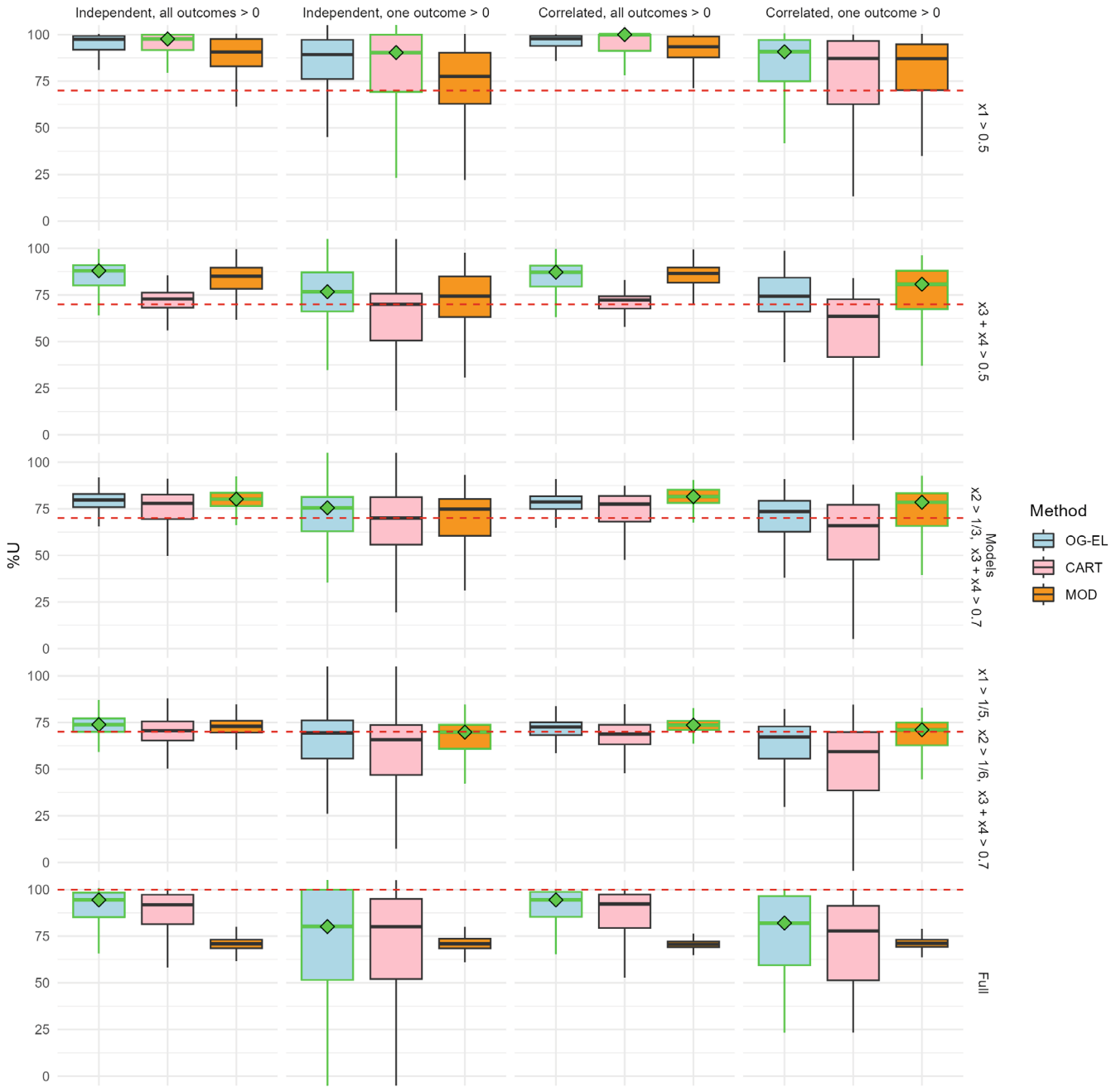

FIGURE 2 The distributions of $\%U$ in the estimated best subgroup from subgroup identification methods across multiple-outcome settings. The method with highest median $\%U$ for each scenario is indicated by green outline and marker. $\%U_{all}$ when no subgroup selection is performed is indicated by a red dashed line for each model (Models 1–4, 70%; Model 5, 100%).

the type I error rate in Model 0. Power was evaluated with the testing procedure described in Section 3.2. using the permuted P-value from $K = 2$ cross-validation folds and 100 permutations. The use of $K = 2$ has been recommended in the literature.[27]

Figure 2 summarizes the $\%U$ in the estimated best subgroup from subgroup identification methods across multiple-outcome settings. Model 0 is not included as $\%U$ under the null hypothesis is not informative. The subgroup estimation accuracy is assessed by medians (the higher the better) and inter-quartile ranges (IQR, the smaller the better) of $\%U$. Overall, in the majority of situations, OG-EL has the best performance in terms of the highest medians that are much better (+20%) than or close to (±5%) the benchmark values $\%U_{all}$ and the smallest IQRs. In situations where OG-EL does not have the highest median, it generally still has comparable performance compared to the best method.

TABLE 1 Feasibility of estimating the weights from data in the definition of the best subgroup, comparing weights estimated from data vs pre-specified weights.

Methods of weighting	Weights estimated from data	Pre-specified weights (1,0,0)	(0,1,0)	(0,0,1)	$\begin{pmatrix}1/\sqrt{3},\\1/\sqrt{3},\\1/\sqrt{3}\end{pmatrix}$	Weights estimated from data	Pre-specified weights (1,0,0)	(0,1,0)	(0,0,1)	$\begin{pmatrix}1/\sqrt{3},\\1/\sqrt{3},\\1/\sqrt{3}\end{pmatrix}$
Model	**Independent outcomes**					**Correlated outcomes**				
All outcomes > 0										
1	97 (92, 99)	92 (83, 98)			97 (92, 99)	98 (94, 99)	93 (84, 97)			98 (95, 99)
2	88 (80, 91)	83 (74, 90)			88 (80, 91)	87 (80, 91)	82 (73, 89)			88 (81, 91)
3	80 (76, 83)	78 (72, 81)			80 (76, 83)	79 (75, 82)	77 (71, 80)			79 (76, 82)
4	74 (70, 77)	72 (66, 76)			74 (70, 77)	73 (68, 75)	70 (64, 74)			73 (70, 75)
5 (Full)	94 (85, 98)	87 (72, 96)			94 (86, 98)	94 (85, 99)	86 (74, 96)			97 (91, 99)
One outcome > 0										
1	89 (76, 97)	90 (76, 97)			78 (47, 93)	90 (75, 97)	90 (76, 97)			81 (60, 93)
2	77 (66, 87)	77 (66, 88)			70 (45, 84)	74 (66, 84)	75 (68, 87)			71 (54, 81)
3	75 (63, 81)	76 (66, 81)			68 (40, 80)	73 (63, 79)	74 (66, 80)			68 (49, 77)
4	69 (56, 76)	70 (57, 75)			61 (35, 74)	67 (56, 73)	68 (59, 73)			62 (45, 71)
5 (Full)	80 (51, 100)	82 (52, 100)			68 (37, 100)	82 (59, 96)	86 (67, 97)			74 (53, 91)

Note: Median (25th percentile, 75th percentile) values of $\%U$ in the estimated best subgroup from OG-EL are reported. When (1,0,0) (0,1,0) (0,0,1) is pre-specified, the best of these vectors is selected as the estimated weight vector. For "all outcomes > 0", $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$ is the optimal weight vector; for "one outcome > 0", the optimal weight vector is one of (1,0,0) (0,1,0), (0,0,1).

It is never notably outperformed by other methods, especially MOD which is favored in Model 14 because it only selects half-spaces which coincide with the true prevalence of S_{true}. Naturally, MOD performs poorly when there is no subgroup with differential treatment effect in the population, and the treatment should be recommended to everyone (Model 5). CART generally has the worst performances, likely because it is a greedy search algorithm with known variable selection bias. All methods exhibit slight drop in performances from "all outcomes > 0" to "one outcome > 0" scenarios (given the same correlation structure). This is as expected since there is a treatment effect only with respect to one of the three outcomes. Correlation among outcomes, "independent" vs "correlated" outcomes, does not affect the quality of subgroup estimation by much within the same outcome scenarios. The corresponding table with numeric results for Figure 2 is in Supplement.

Table 1 summarizes the distributions of $\%U$ in the estimated best subgroup from OG-EL across multiple-outcome settings to compare our proposed approach of estimating the weights ("estimated weights") in the definition of the best subgroup with using pre-specified weights. For pre-specified weights we considered a subgroup estimation method where we know the correct set of weights $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$. The vector $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$ is the optimal weight vector for the "all outcomes > 0" scenarios. For the "one outcome > 0" scenarios, we know that one outcome has a treatment effect greater than 0 but we do not know which one. In subgroup estimation, we choose one of the three vectors (1,0,0), (0,1,0), and (0,0,1). Table 1 shows that it is feasible to estimate the weights while the subgroup is being estimated. It yields $\%U$ similar (within 5%) to when the optimal weight vector is used.

Table 2 summarizes the simulation results of variable selection accuracy (%) in the estimated best subgroup with estimated weights from subgroup identification methods across multiple-outcome settings. The variable selection accuracy is assessed with % of runs with $\widehat{S}$ that has the exact set of biomarkers in S_{true}. In Model 1 and 2, OG-EL has the best performances in terms of the highest selection rates across situations, up to ~80% accuracy in Model 1, and outperforms other methods by large margins (up to +70%) as well. As S_{true} becomes more complex (Model 3, 4, and 5), MOD performs notably better (up to +40%) than OG-EL. CART performs the worst in all situations (especially Model 5) as expected due to its known variable selection bias.

Table 3 summarizes simulation results of empirical powers from the described testing procedure in Section 3.2. in the estimated best subgroup from OG-EL across multiple-outcome settings, at 0.05 two-sided significance levels, comparing

TABLE 2 Variable selection accuracy (%) in the estimated best subgroup from subgroup identification methods across multiple-outcome settings.

Methods of analysis	OG-EL	CART	MOD	OG-EL	CART	MOD
Model	**Independent outcomes**			**Correlated outcomes**		
All outcomes > 0						
1	80	61	18	83	60	21
2	62	21	27	65	21	24
3	14	6	29	15	6	31
4	31	2	55	27	2	71
5 (Full)	12	1	30	12	4	30
One outcome > 0						
1	56	44	9	62	39	12
2	34	14	15	36	11	17
3	8	4	16	10	6	24
4	20	3	42	19	5	59
5 (Full)	16	4	25	12	3	30

TABLE 3 Power in the estimated best subgroup from OG-EL from the 0.05 two-sided significance test.

Methods of weighting	Weights estimated from data	Pre-specified weights (1,0,0) (0,1,0) (0,0,1)	Pre-specified weights $\left(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)$	Weights estimated from data	Pre-specified weights (1,0,0) (0,1,0) (0,0,1)	Pre-specified weights $\left(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)$
Model	**Independent outcomes**			**Correlated outcomes**		
0 (Null)	0.048	0.047	0.049	0.050	0.049	0.050
All outcomes > 0						
1	0.98	0.77	0.99	0.96	0.81	1.00
2	0.97	0.72	0.98	0.93	0.73	0.99
3	0.95	0.67	0.98	0.92	0.68	0.99
4	0.93	0.60	0.96	0.87	0.63	0.98
5 (Full)	0.88	0.53	0.93	0.81	0.56	0.96
One outcome > 0						
1	0.64	0.64	0.55	0.60	0.65	0.52
2	0.59	0.59	0.49	0.55	0.59	0.50
3	0.57	0.57	0.50	0.52	0.54	0.47
4	0.55	0.55	0.47	0.49	0.54	0.44
5 (Full)	0.45	0.47	0.45	0.43	0.45	0.40

Note: For Model 0 (Null) the result is the estimated type I error rate. When (1,0,0) (0,1,0) (0,0,1) is pre-specified, the best of these vectors is selected as the estimated weight vector. For "all outcomes > 0", $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$is the optimal weight vector; for "one outcome > 0", the optimal weight vector is one of (1,0,0) (0,1,0), (0,0,1).

different weighting approaches. The empirical power is calculated as the percentage of trials with significant *P*-values. Under the null hypothesis (Model 0), doing permutation preserves the overall type I error rate at the nominal level (0.05). When there are three possible weight vectors (1,0,0), (0,1,0), and (0,0,1), the Bonferroni approach is used to compute the adjusted *P*-value. The empirical powers mostly correspond to subgroup estimation accuracy results in Table 1: the better the subgroup estimation, the higher the empirical power. Using optimal weights is the most powerful. Estimating weights is comparable in power to using optimal weights and better than using wrong fixed weights. Knowing the corrects weights is more beneficial when outcomes are correlated probably due to reduced information contained in correlated outcomes compared to independent outcomes.

For brevity, for Tables 1 and 3, we only present results from OG-EL because it has the best overall performance. For sensitivity analyses, we investigated multiple-outcome settings with the same outcome types (eg, three continuous outcomes with a total of six responses generated from independent or correlated multivariate Gaussian distributions). We did not see any noticeable difference in results when outcomes were of the same type. We also investigated cross-validation performances and did not see any noticeable difference for values of K between 2 and 10. In addition, we repeated all settings with correlated biomarkers. We used a multivariate Gaussian copula to achieve a correlation of 0.2 between X_1 and X_2, a correlation of 0.8 between X_3 and X_4, and a correlation of 0 between other pairs of biomarkers. Correlated biomarkers result in a slight drop in performance across the board. However, the conclusions align with those from settings with independent biomarkers.

5 | RIVUR APPLICATION

In the RIVUR[32,33] trial, a total of 607 children 2–71 months of age were randomized in the ratio of 1:1 to receive either an antibiotic prophylaxis or placebo daily for 2 years. They were followed to ascertain the primary outcome of UTI recurrence with renal scarring being an important secondary outcome. The primary analysis by RIVUR investigators[42] found that long-term antimicrobial prophylaxis substantially reduced the risk of UTI recurrences by 50%, comparing the prophylaxis and placebo groups. Not enough renal scarring events were observed to make a conclusion about the effect of prophylactic antibiotic on scarring. We applied our proposed definition of best subgroup to estimate the best subgroups of children, defined by three baseline biomarkers: age, VUR, and bowel-bladder dysfunction (BBD). Because use of long-term antimicrobial prophylaxis may lead to the development of antibiotic resistance and alterations of microbiome, and because the number needed to treat observed in the RIVUR trial was relatively large, there is an interest in identifying higher-risk subgroups of children that would benefit the most from long-term antimicrobial prophylaxis.

Details on biomarkers considered are in Supplement. We analyzed two binary outcomes: UTI recurrence and occurrence of renal scarring. We adopted a complete-case approach by omitting children with any missing outcomes or biomarkers, so the resulting RIVUR sample consisted of 444 children, 224 in the placebo control arm ($T = 0$) and 220 in the antimicrobial prophylaxis treatment arm ($T = 1$). The treatment effect (treatment – control) of UTI recurrence is 59/224-27/220 = 0.141, with one-sided *P*-value <.001, while that for occurrence of renal scarring is 19/224-18/220 = 0.003, with one-sided *P*-value = .46.

We applied the proposed definition of best subgroup with estimated weights to the two outcomes and estimated the best subgroup from the entire sample with the pool of candidates generated from OG-EL, considering only subgroups with at least 10% prevalence. We computed the adjusted *P*-value of testing the treatment effect of each outcome independently in the estimated subgroup, from the described 2-fold cross-validation procedure with 1000 permutations. We then calculated an ad-hoc honest estimate of risk difference (details are in Supplement). As comparisons, we analyzed the single-outcome setting of UTI recurrence and calculated adjusted *P*-value and risk difference accordingly. Results are presented in Table 4.

Under the single-outcome setting, UTI recurrence yielded an estimated best subgroup which consists of all children except those without BBD (96% of the RIVUR sample), only slightly smaller than the whole sample. Under the two-outcome setting, the estimated optimal weights were $\sqrt{0.95}$ for UTI recurrence and $\sqrt{0.05}$ for renal scarring, with the estimated best subgroup yielding 19% of the RIVUR sample with much larger treatment effect (for UTI recurrence) than the one in the subgroup estimated from UTI recurrence only. Our proposed approach allowed incorporating the information from renal scarring into that of UTI recurrence to identify a subgroup with significantly larger clinical benefit from long-term antimicrobial prophylaxis.

TABLE 4 The estimated best subgroup in the RIVUR trial from two outcomes and from a single outcome.

Outcomes	Subgroup composition	Size (%)	UTI recurrence		Renal scarring	
			Risk difference	*P*-value	Risk difference	*P*-value
UTI & Renal scarring	Grade II VUR & BBD present & any index UTI Grade III VUR & BBD present & febrile index UTI Grade III VUR & any BBD & symptomatic index UTI Grade IV VUR	85 (19%)	0.24	.02	0.006	.95
UTI	All except BBD absent & symptomatic index UTI	427 (96%)	0.10	.01	0.01	.72

Note: Risk difference and the *P*-value are from a cross-validation procedure.

6 | CONCLUSION

We have expanded the definition of the best subgroup to multiple outcome setting. Our proposed definition offers a trade-off between the subgroup size and the CATE in the subgroup, with respect to each of the outcomes. A part of finding the best subgroup is finding the best set of weights that reflects the relative treatment effect measured with respect to each of the outcomes. These weights give the best power if the treatment effect is tested based on the weighted combination of the test statistics computed for each of the outcomes. In cases where there is no pre-existing knowledge about the relative importance of the outcomes, we suggest determining the weights from the data while identifying the optimal subgroup. Our work was motivated by the PrecISE clinical trial in severe asthma. Three primary outcomes were selected in asthma control, symptoms, and lung function domains. Some existing asthma medications only work in one or two domains, not all. This was the reason for selecting all three as primary endpoints and the reason for estimating the weights from data obtained in this trial. Conversely, if there is pre-existing knowledge, one option is to employ a set of clinically appropriate, pre-specified weights that assign higher weights to more important outcomes. Alternatively, information like time to the first event (if available for all outcomes) could be utilized to construct a composite win ratio as the measure of utility to impose hierarchical constraints to the outcomes.[43]

Our definition can be used in two main contexts. The first is a post-hoc analysis to identify the best subgroup. Due to a complex and unknown correlation structure among the test statistics corresponding to multiple outcomes, resampling methods are currently the most widely used and validated approaches to obtain an unbiased estimate of the CATE in the subgroup. The second setting is in a multi-stage randomized clinical trial with prospective enrichment. Using an example of a two-stage trial, after the first stage, the best subgroup and its associated weights are identified. The Stage 1 *P*-value to test the CATE of the estimated subgroup with respect to the optimal set of weights is adjusted using cross-validation with permutation. Stage 2 population is enriched and only participants from the subgroup are enrolled. The optimal weights estimated in Stage 1 are used to combine outcome-specific Stage 2 test statistics. The Stage 1 and Stage 2 *P*-values are then combined to obtain a trial-wise *P*-value.

ACKNOWLEDGEMENTS

Work of Anastasia Ivanova is supported in part by the National Heart, Lung, and Blood Institute, National Institutes of Health, grant U24 HL138998.

DATA AVAILABILITY STATEMENT

Data subject to third party restrictions: The data that support the findings of this study are available from National Institutes of Health (NIH). Restrictions apply to the availability of these data, which were used under license for this study.

ORCID

Beibo Zhao https://orcid.org/0000-0001-8355-7143
Anastasia Ivanova https://orcid.org/0000-0003-4321-2073

REFERENCES

1. Lipkovich I, Dmitrienko A, D' Agostino SBR. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*. 2017;36(1):136-196.

2. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat*. 2013;7(1):443-470. 428.
3. Huber C, Benda N, Friede T. A comparison of subgroup identification methods in clinical drug development: simulation study and regulatory considerations. *Pharm Stat*. 2019;18(5):600-626.
4. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30(24):2867-2880.
5. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc*. 2014;109(508):1517-1532.
6. Lai TL, Lavori PW, Liao OY-W. Adaptive choice of patient subgroup for comparing two treatments. *Contemp Clin Trials*. 2014;39(2):191-200.
7. Lai TL, Lavori PW, Tsang KW. Adaptive enrichment designs for confirmatory trials. *Stat Med*. 2019;38(4):613-624.
8. Joshi N, Fine J, Chu R, Ivanova A. Estimating the subgroup and testing for treatment effect in a post-hoc analysis of a clinical trial with a biomarker. *J Biopharm Stat*. 2019;29(4):685-695.
9. Joshi N, Nguyen C, Ivanova A. Multi-stage adaptive enrichment trial design with subgroup estimation. *J Biopharm Stat*. 2020;30(6):1038-1049.
10. Chen G, Zhong H, Belousov A, Devanarayan V. A prim approach to predictive-signature development for patient stratification. *Stat Med*. 2015;34(2):317-342.
11. Huang X, Sun Y, Trow P, et al. Patient subgroup identification for clinical drug development. *Stat Med*. 2017;36(9):1414-1428.
12. Zhang Z, Li M, Lin M, Soon G, Greene T, Shen C. Subgroup selection in adaptive signature designs of confirmatory clinical trials. *J R Stat Soc Ser C Appl Stat*. 2017;66(2):345-361.
13. Wang T, Wang X, George SL, Zhou H. Design and analysis of biomarker-integrated clinical trials with adaptive threshold detection and flexible patient enrichment. *J Biopharm Stat*. 2020;30(6):1060-1076.
14. Su X, Tsai C-L, Wang H, Nickerson D, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10:141-158.
15. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. *The. Int J Biostat*. 2008;4(1):1-26.
16. Seibold H, Zeileis A, Hothorn T. Model-based recursive partitioning for subgroup analyses. *Int J Biostat*. 2016;12(1):45-63.
17. Seibold H, Zeileis A, Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Stat Methods Med Res*. 2018;27(10):3104-3125.
18. Xu Y, Yu M, Zhao Y-Q, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*. 2015;71(3):645-653.
19. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*. 2017;73(4):1199-1209.
20. Ivanova A, Israel E, Lavange LM, et al. The precision interventions for severe and/or exacerbation-prone asthma (precise) adaptive platform trial: statistical considerations. *J Biopharm Stat*. 2020;30(6):1026-1037.
21. Fuhlbrigge AL, Bengtsson T, Peterson S, et al. A novel endpoint for exacerbations in asthma to accelerate clinical development: a post-hoc analysis of randomised controlled trials. *Lancet Respir Med*. 2017;5(7):577-590.
22. Keren R. Pediatrics Rivur trial introduction. *Pediatrics*. 2008;122(Suppl 5):S231-S232.
23. Zhao B, Ivanova A, Shaikh N. Antimicrobial prophylaxis for vesicoureteral reflux: which subgroups of children benefit the most? *Pediatr Nephrol*. 2024.
24. Loh W-Y, Zheng W. Regression trees for longitudinal and multiresponse data. *Ann Appl Stat*. 2013;7(1):495-522.
25. Loh W-Y, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med*. 2015;34(11):1818-1833.
26. Zhang P, Liu P, Ma J, Shentu Y. Value function guided subgroup identification via gradient tree boosting: a framework to handle multiple outcomes for optimal treatment recommendation. *Stat Biopharmaceut Res*. 2022;14(4):523-531.
27. Talisa VB, Chang CH. Learning and confirming a class of treatment responders in clinical trials. *Stat Med*. 2021;40(22):4872-4889.
28. Hastie T, Friedman J, Tibshirani R. Linear methods for regression. In: Hastie T, Friedman J, Tibshirani R, eds. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York; 2001:41-78.
29. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1984.
30. LeBlanc M, Crowley J. Relative risk trees for censored survival data. *Biometrics*. 1992;48(2):411-425.
31. Stallard N. Adaptive enrichment designs with a continuous biomarker. *Biometrics*. 2023;79(1):9-19.
32. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(1):99-119.
33. Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clin Cancer Res*. 2010;16(2):691-698.
34. Guo X, He X. Inference on selected subgroups in clinical trials. *J Am Stat Assoc*. 2021;116(535):1498-1506.
35. Zhang Z, Chen R, Soon G, Zhang H. Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. *Stat Med*. 2018;37(1):1-11.
36. Zhao B, Ivanova A, Fine J. Inference on subgroups identified based on a heterogeneous treatment effect in a post hoc analysis of a clinical trial. *Clin Trials*. 2023;20(4):370-379.
37. Simon R. Development and validation of biomarker classifiers for treatment selection. *J Stat Plan Inference*. 2008;138(2):308-320.
38. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res*. 2005;11(21):7872-7878.
39. Whitlock MC. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J Evol Biol*. 2005;18(5):1368-1373.

40. Sklar A. Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture Notes-Monograph Series*. 1996;28:1-14.
41. *R: A Language and Environment for Statistical Computing [Computer Program]*. Vienna, Austria: R Foundation for Statistical Computing; 2022.
42. Hoberman A, Greenfield SP, Mattoo TK, et al. Antimicrobial prophylaxis for children with vesicoureteral reflux. *N Engl J Med*. 2014;370(25):2367-2376.
43. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33(2):176-182.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhao B, Fine J, Ivanova A. Finding the best subgroup with differential treatment effect with multiple outcomes. *Statistics in Medicine*. 2024;43(13):2487-2500. doi: 10.1002/sim.10083