Identifying Heterogeneous Treatment Effects in Multiple Outcomes using Joint Confidence Intervals

Peniel N. Argaw[1*] PENIEL@G.HARVARD.EDU
Elizabeth Healey[2,3*] EHEALEY@MIT.EDU
Isaac S. Kohane[3] ISAAC_KOHANE@HMS.HARVARD.EDU
[1] *Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA*
[2] *Massachusetts Institute of Technology, Cambridge, MA*
[3] *Harvard Medical School, Boston, MA*

Abstract

Heterogeneous treatment effects (HTEs) are commonly identified during randomized controlled trials (RCTs). Identifying subgroups of patients with similar treatment effects is of high interest in clinical research to advance precision medicine. Often, multiple clinical outcomes are measured during an RCT, each having a potentially heterogeneous effect. Recently there has been high interest in identifying subgroups from HTEs, however, there has been less focus on developing tools in settings where there are multiple outcomes. In this work, we propose a framework for partitioning the covariate space to identify subgroups across multiple outcomes based on the joint CIs. We test our algorithm on synthetic and semi-synthetic data where there are two outcomes, and demonstrate that our algorithm is able to capture the HTE in both outcomes simultaneously.

Keywords: causal inference, heterogeneous treatment effect, joint CIs, subgroup analysis

1. Introduction

The goal of clinical trials is to evaluate the efficacy of interventions, often by comparing the outcomes between treatment and control therapies. The efficacy is assessed through estimating the treatment effect. HTEs can explain the variability of treatment effects in a population over a covariate space by defining a set of subgroups with similar treatment effects, and these subgroups can then be analyzed in ways to advance precision medicine (Varadhan et al., 2013). For example, subgroup analysis can provide insight about which types of patients may respond exceptionally well or poorly to a given therapy (Rekkas et al., 2020).

In many cases, researchers and clinicians are interested in a given therapy's treatment effect on multiple outcomes. While there has been significant interest in developing techniques to discover HTEs from RCT data, they primarily focus on settings where there is just one outcome of interest. These do not capture the complexities of real-world scenarios where the therapy causes multiple effects (Berkey et al., 1996). Certain chronic diseases call for management through treatments that affect multiple clinical endpoints. Clinical trials often evaluate the treatment effect on primary and secondary outcomes, and it is common for clinical trials

[*] Equal contribution

to have multiple primary outcomes (Vickerstaff et al., 2015).

Clinicians, researchers, and pharmaceutical companies can benefit from knowing subgroup characteristics across multiple endpoints. For instance, a certain medication may show high efficacy in two subgroups, but show adverse side effects in one of the groups. Hence, it is critical that computational research focuses attention on the development of methods to analyze RCT data that captures multiple clinical endpoints.

One of the challenges in studying HTEs on multiple different outcomes is ensuring robust multivariate treatment effect estimates among subgroups. In the single outcome setting, recent work has explored subgroup identification in order to optimally separate treatment effects (Rekkas et al., 2020), and others identified subgroups by optimizing for intra- and inter-group variation and confidence intervals (Lee et al., 2020). Algorithms that identify subgroups without accounting for intra-group variation run the risk of having subgroups with large CIs. This depreciates the robustness of the treatment effect estimate within a group and can limit the clinical utility of the findings. This can be further complicated when evaluating on multiple outcomes.

In our work, we offer a solution to this problem by proposing a novel method for partitioning treatment effects using the joint CIs of multiple outcome variables. We extend upon the partitioning framework from Lee et al. (2020) by generating joint CIs using conformal prediction and quantile regression (Lei et al., 2018; Romano et al., 2019; Lei and Candès, 2021). We evaluate our approach on synthetic and semi-synthetic datasets inspired by clinical data and generalize our algorithm for multiple outcomes (specifically two outcomes in this study). We refer to our method as Multiple Outcome Partitioning using Joint Confidence Intervals, MOP-JCI.

Our Key Contributions

1. We extend upon a framework for partitioning the covariate space to identify a set of subgroups with similar treatment effects across multiple outcomes using joint CIs.

2. We deploy and evaluate a quantile individual treatment effect (ITE) estimator in the partitioning algorithm.

3. We evaluate our approach on synthetic and semi-synthetic RCT datasets and show the robustness of our method on datasets containing correlation and heteroskedasticity.

2. Related Work

2.1. Subgroup Analyses

Subgroup analysis is a common approach to identifying heterogeneities in treatment effect. Methodologies to estimate the heterogeneity include statistical tests (Assmann et al., 2000; Alosh et al., 2015), Bayesian modeling (Jones et al., 2011; Pennello and Rothmann, 2018) and recursive partitioning (Su et al., 2009; Athey and Imbens, 2016; Lee et al., 2020; Seibold et al., 2016). Lee et al. (2020) proposed a confidence criterion for use in recursive partitioning, derived from the CIs of any mean ITE estimator, to ensure homogeneity within subgroups. These approaches, however, are all limited to a single outcome variable.

2.2. Multiple Outcomes

Multiple outcomes are common in RCTs and recent work has considered analyzing treatment effects in the setting of multiple outcomes. Kennedy et al. (2019) presented an approach to estimate the effects of multiple

outcomes using a common scale, and discussed the dependency of treatment effects on covariates. Wu et al. (2022) proposed a personalized policy generation method in the setting of multiple outcomes by weighting the treatment effect of each outcome. Yao et al. (2022) proposed a method of treatment effect estimation that utilizes data across multiple outcomes. Yoon et al. (2011) evaluates likelihood-based methods to jointly test treatment effects across multiple outcomes. These approaches, however, do not focus on identifying subgroups where each group of patients show similar characteristics across all outcomes.

3. Methods

We begin by defining a preliminary framework which we use to estimate treatment effect. Next, we discuss the CI generation techniques used in our implementation. We then describe joint CI estimation for multiple outcomes and show the integration into a recursive partitioning algorithm to construct subgroups.

3.1. Preliminaries

We consider a setup of a RCT where there are two different outcomes of interest, outcome A and outcome B. Namely, we have a total of N samples each with covariates X_i, treatment assignment t_i, and outcome variables Y_i^A and Y_i^B for $i = 1, ..., N$. Here, t_i is a binary variable $\{0, 1\}$ representing the treatment group assignment for the sample. The outcome variables, Y_i^A and Y_i^B, are scalar, continuous values for outcomes A and B, respectively. Our goal is to determine the ITE for each outcome as defined by $\mathbb{E}[Y_i^A(1) - Y_i^A(0)|X = x]$ and $\mathbb{E}[Y_i^B(1) - Y_i^B(0)|X = x]$ where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes for each sample i had they been treated with 1 or 0, respectively.

The ITE estimate for each outcome is the difference of the two regression models for the control and treated, $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ respectively. The regression models for outcomes A and B are defined as $\hat{\mu}_0^A(x) = \mathbb{E}[Y^A(0)|X = x]$, $\hat{\mu}_1^A(x) = \mathbb{E}[Y^A(1)|X = x]$, $\hat{\mu}_0^B(x) = \mathbb{E}[Y^B(0)|X = x]$, $\hat{\mu}_1^B(x) = \mathbb{E}[Y^B(1)|X = x]$.

CIs can be generated from each of the regressors using split conformal regression (SCR) (Lei et al., 2018). SCR initially splits the samples into two equal-sized sets, a training set I_{tr} and a validation set I_{val}, then trains a regressor $\hat{\mu}^{I_{tr}}$, calculates the residuals of the trained model on the validation set I_{val}. The resulting CI bounds can be defined as $\hat{C}(x) = \left[\hat{\mu}^{I_{tr}}(x) - \hat{Q}_{1-\alpha}^{I_{val}} \ , \ \hat{\mu}^{I_{tr}}(x) + \hat{Q}_{1-\alpha}^{I_{val}}\right]$ where $\hat{Q}_{1-\alpha}^{I_{val}}$ is the $(1-\alpha)(1+\frac{1}{|I_{val}|})$-th quantile of the residuals $\{|y_i - \hat{\mu}^{I_{tr}}(x_i)|\}_{i \in I_{val}}$, and α is the miscoverage rate used to ensure a coverage guarantee for each outcome y such that $\mathbb{P}[y \in \hat{C}] \geq 1 - \alpha$.

3.2. Split conformal quantile regression

Split conformal quantile regression (SCQR) is an alternative approach to estimate the ITE CIs for each regression model, $\hat{\mu}_0(X)$ and $\hat{\mu}_1(X)$, as described in Romano et al. (2019). Again, we first split the data equally into a training and validation set, I_{tr} and I_{val}. The training set is used to fit the two quantile regression models, $\hat{q}_\alpha^{hi}$ and $\hat{q}_\alpha^{low}$ for a miscoverage rate α. Using these estimators, we compute the calibration scores E_i for each $i \in I_{val}$ by $E_i = \max\{\hat{q}_\alpha^{low}(x_i) - Y_i, Y_i - \hat{q}_\alpha^{hi}(x_i)\}$. The CIs of the estimator are $\hat{C}(x) = \left[\hat{q}_\alpha^{low}(x) - \hat{Q}_{1-\alpha}^{I_{val}}(E) \ , \hat{q}_\alpha^{hi}(x) + \hat{Q}_{1-\alpha}^{I_{val}}(E)\right]$, where $\hat{Q}_{1-\alpha}^{I_{val}}$ is the $(1 - \alpha)(1 + \frac{1}{|I_{val}|})$-th quantile of $\{|E_i|\}_{i \in I_{val}}$.

3.3. Joint CIs for ITE estimate

We calculate the conformalized ITE intervals by jointly considering the treated regression and control regression. We use the naive approach outlined in Lei and Candès (2021) which directly compares the two intervals adjusts the miscoverage rate by dividing α by 2. Accordingly, we define the CIs for the treated population as $[\hat{C}^{low}_{\alpha/2}(1;x)\ ,\ \hat{C}^{hi}_{\alpha/2}(1;x)]$ and for control population as $[\hat{C}^{low}_{\alpha/2}(0;x), \hat{C}^{hi}_{\alpha/2}(0;x)]$ each with a coverage of $1-\alpha/2$. The CI of the ITE estimator is defined as $\hat{C}_{ITE}(x) = [\hat{C}^{low}_{\alpha/2}(1;x) - \hat{C}^{hi}_{\alpha/2}(0;x),\ \hat{C}^{hi}_{\alpha/2}(1;x) - \hat{C}^{low}_{\alpha/2}(0;x)]$.

3.4. Joint CIs for multiple outcomes

In the single outcome case, coverage is guaranteed for each outcome y such that $\mathbb{P}[y \in \hat{C}] \geq 1-\alpha$, where α was the miscoverage rate of the ITE estimate. We apply the Bonferroni correction to our coverage term in order to adjust the CIs for each outcome and guarantee a specified overall coverage across all the outcomes. This is done by taking the joint probability that each outcome's CIs are within a given coverage. Concretely, we divide the miscoverage rate by d, where d is the total number of outcomes. Combining this adjustment with the previous adjustment in section 3.3, we set the miscoverage rate as $\frac{\alpha}{2d}$ for each treated and control regressor, thus ensuring $1-\alpha$ coverage across the treatment effect of all outcomes ($\mathbb{P}[y \in \hat{C}_{ITE}] \geq 1-\frac{\alpha}{2d}$).

3.5. Recursive Partitioning Algorithm

We build upon the robust recursive partitioning algorithm (R2P) proposed Lee et al. (2020) to partition the data into subgroups based on the covariate space. We adapt their confidence criterion for use in the setting of multiple outcomes. They define a confidence criterion that aims to maximize heterogeneity across the subgroups and maximize homogeneity within the subgroups. They do this by minimizing the expected CI width W_g, with the expected absolute deviation V_g within a group g. The expected width W_g is defined as $\mathbb{E}[|\hat{C}_g(x)|]$, and the deviation $V_g = \mathbb{E}[\hat{v}_g(x)]$ where $v_g = (\hat{\mu}^{mean}_g - \hat{\mu}^{up}_l(x))\ \mathbb{I}\left[\hat{\mu}^{mean}_l > \hat{\mu}^{up}_l(x)\right] + (\hat{\mu}^{low}_g(x) - \hat{\mu}^{mean}_g)\ \mathbb{I}\left[\hat{\mu}^{mean}_l < \hat{\mu}^{low}_g(x)\right]$.

The expected absolute deviation V_g within a group g can otherwise be explained as the error between the the CI bound and the average outcome. Together, the partitioning is done with the following objective.

$$\text{minimize} \sum_{g \in \Pi} \lambda W_g + (1-\lambda) V_g$$

where λ is a hyperparameter used to vary the weight on V_g and W_g and Π is the set of partitions. We extend this criterion to work with multiple outcomes by summing the regions for each outcome, using predefined weights. In the two outcome case, the objective function is as follows:

$$\text{minimize} \sum_{g \in \Pi} \lambda(\beta W^A_g + (1-\beta) W^B_g) + (1-\lambda)(\beta V^A_g + (1-\beta) V^B_g)$$

Here, β is a tuning parameter to weight the outcomes. It can be tuned to give preference for one outcome over another, and to account for differences in expected magnitude of the outcomes. We provide a modified objective function for more than two outcomes in Appendix C.

For the SCQR method, we simplify the objective function to the setting where λ is 0. This is because in the SCQR setting, the CIs for each covariate are determined by the quantile estimator. They do not change with further calibration after each split. Thus, the objective function for the SCQR method in the setting of two outcomes is defined as:

$$\text{minimize} \sum_{g \in \Pi} (\beta V^A_g + (1-\beta) V^B_g)$$

We further adapt the robust recursive partitioning algorithm proposed in Lee et al. (2020) to partition on two outcomes and to work with a quantile estimator in Algorithm 1 (the partitioning algorithm for SCR can be found in the Appendix C).

Algorithm 1: SCQR Recursive Partitioning on Two Outcomes

Input: G_{node}, data
for *Covariate j in data* **do**
  for *Unique value x of covariate j* **do**
    Split data in two branches on x
    Compute V^A for each branch
    Compute V^B for each branch
    Set G_{split} to G_{node} less the sum of $(\beta V^A + (1-\beta)V^B)$ from each branch
  end
end
Save best G_{split} with covariate j
if $G_{split} > \gamma G_{node}$ **then**
  $G_{node} \leftarrow G_{split}$
  Partition on each branch
else
  Set current node to leaf
end

To work with the quantile estimator, we first perform SCQR on the training data using the I_{tr} to fit the random forest quantile estimator and I_{val} to compute the confidence metrics. We then compute $\hat{C}^A_{ITE}$ and $\hat{C}^B_{ITE}$ from I_{val} of each outcome. Since the SCQR estimates the quantile distribution of the treatment effect across the covariate space, there is no need for recalibrating during the partitioning, reducing computational burden.

We use the I_{val} data to partition the data using a recursive function. We start with the entire set I_{val} as the root node and use Algorithm 1 to create nodes by computing the a joint confidence score using V_g. The recursive function takes in a node and the $\hat{C}_{ITE}$ of each estimator to calculate the V_g. The best gain is computed for each split along a single covariate as G_{split}. A node is split into branches when the candidate split, G_{split}, is greater than γG_{node}, where γ controls for regularization of the number of subgroups. The resulting leaves make up the subgroups.

4. Experimental Design

4.1. Experiments

We generate a set of experiments to evaluate our method and the use of the two different conformal regression techniques.

1. Baseline (R2P): We use R2P to partition on a single outcome at a time. We observe the subgroup formation from partitioning on each outcome separately.

2. Our method (MOP-JCI)

   - SCR approach: We use SCR to generate joint CIs using the Bonferroni correction to guarantee joint coverage of both outcomes. We partition using the CI regions of both outcomes in the minimization function.
   - SCQR approach: We use SCQR to generate joint CIs using the Bonferroni correction to guarantee joint coverage of both outcomes. We partition using the CI regions of both outcomes in the minimization function.

Moreover, we experimented on various ITE estimators. These estimators included Causal Multi-task Gaussian Process (CMGP) (Alaa and van der Schaar, 2017), Random Forest (RF), and Quantile Random Forest (QRF). For CMGP, we use the implementation provided by the authors. RF and QRF are all adapted from *scikit-learn* and *scikit-garden* to estimate confi-

dence bounds across the treated and control populations for multiple outcomes. See Appendix B and C for details on hyperparameter tuning. Additionally, this study is based on the assumptions listed in Appendix C. Our implementation is available at https://github.com/pargaw/MOP-JCI.

4.2. Evaluation Metrics

We evaluate the statistical significance of the defined subgroups and the precision of the regressors. As our goal was to maximize heterogeneity across groups and homogeneity within groups, we evaluated the variance found within and across the groups. Variance across the groups was defined as $V_{across} = Var(Mean(S_g^{test})_{g=1}^{G})$ where S_g^{test} is the set of test samples in group g and G is the total number of subgroups. Variance within the groups was defined as $V_{within}(S^{test}) = \frac{1}{G}\sum_{g=1}^{G} Var(S_g^{test})$. Further, we evaluate the true coverage of the CIs generated (Cov) by computing the percentage of time that both outcome predictions fall within the CI with miscoverage set at 0.1. We report the mean width of the CIs (CI Width) as well. Additionally, we evaluated the error of our ITE estimators using the precision in estimation of heterogeneous effect (PEHE) (Hill, 2011).

In our experiments, we run each algorithm 30 times to take the mean and standard deviation of the metrics. We additionally compute the percentage of iterations that the partitioning algorithm split the subgroups on the expected covariates (Split Acc). Similarly, we compute the percentage of iterations where the algorithm split on unexpected covariates, (Split Err). Unexpected covariates are covariates that have little or no underlying effect on the outcome distribution. These metrics are added to ensure the subgroups are formed based on ground truth knowledge from the data generation.

4.3. Datasets

We evaluate our work on synthetic and semi-synthetic datasets, where each have two outcomes. Additionally, to assess the robustness of our approach, we evaluate our model on variations of our synthetic dataset that exhibit uncorrelated covariates, correlated covariates, and heteroscedasticity. More details on the outcome generation and distributions for each of the datasets can be found in Appendix A.

4.3.1. Synthetic Data

We adapt synthetic data proposed in Lee et al. (2020) to represent multiple outcomes. The synthetic data was inspired by the clinical trial of remdesivir on COVID-19 (Beigel et al., 2020). The synthetic data consists of simulated versions of covariates used in the trial, with values randomized on varying distributions (see Appendix A). The outcome in the synthetic data is days to clinical improvement, with data simulated to show the relationship between faster clinical improvement and shorter time from symptom onset to starting the trial. We extend the synthetic data to include a second outcome reflective of an adverse event in the trial: alanine aminotransferase (ALT) levels on the last day of the trial (day 28). We focus on the version of this data where the outcomes are uncorrelated. Research has suggested that high ALT levels at baseline may put patients at increased risk for liver function deterioration from remdesivir (Charan et al., 2021). We simulate increased risk of higher end-point ALT as a function of baseline ALT. We simulate data from a trial where the time to improvement is measured as efficacy, and liver function deterioration is measured as an adverse event. In the primary version of this synthetic data, the outcomes are uncorrelated. Additional variations of this dataset

are described and evaluated in Appendix A and D.

4.3.2. Semi-synthetic Data

For the semi-synthetic case, we used the Infant Health Development Program (IHDP) dataset. The dataset was inspired by a real RCT where the goal was to evaluate the efficacy of early intervention to improve the health and development of low-birth-weight, premature infants (Gross, 1993). The deidentified covariates were extracted from the original study, and the outcomes were simulated using the Response Surface B described in Hill (2011). The dataset was adapted for a two-outcome setting by choosing a different covariate to relate to each outcome (Neonatal health index (nnhealth) and mom age respectively). These outcomes and covariate relationships are inspired by the results found in Baumeister and Bacharach (1996).

5. Results

5.1. Synthetic Data

In this section we show the results of the partitioning methods performed on the synthetic dataset. Table 1 shows the baseline partitioning from R2P on a single outcome, and the MOP-JCI methods using the SCR and SCQR approach. Looking at the baseline R2P, when the partitioning is done on outcome 1, clear subgroups are identified with respect to outcome 1 noting that variance in each subgroup is low and variance across each subgroup is high. However, the characteristics of outcome 2 in the corresponding subgroups formed when only outcome 1 is partitioned are not well defined. Whereas, the MOP-JCI methods are able to capture low within group variance and high across group variance for both outcomes simultaneously in the same partition. All MOP-JCI methods are able to identify the expected covariates to split on (denoted by Split Acc). Figure 1 shows the subgroups formed by the MOP-JCI methods. We can see that in both of these cases, the subgroups are well-defined across both outcomes. See Figure S9 for the subgroup formation from R2P on a single outcome and Appendix E for the full subgroup characteristics. The performance metrics at different values of the tuning parameter β are found in Figure S7.

5.2. Semi-synthetic Data

Similarly in the IHDP data, we show the effects of partitioning separately on each outcome using R2P and jointly partitioning using SCR and SCQR using MOP-JCI. Table 2 shows the results from the partitioning algorithms. See Appendix Figure S10 and Figure S11 for the treatment effect of subgroup formations. The MOP-JCI methods are able to achieve similar variance across groups and similar variance within groups for each outcome as when they are partitioned separately. The error metrics when reporting the results of the IHDP data are rather high. This is due to the fact that the generation of the outcomes involves many covariates with a small effect. The accuracy metric is more useful here, since there is one dominant covariate contributing to each outcome effect. The performance metrics at different values of the tuning parameter β are found in Figure S8.

6. Conclusion

With RCT data being widely available, methods to properly analyze the data are essential in order to advance precision medicine. Our work introduces a method that can identify subgroups of patients in an RCT whose response across multiple outcomes is homogeneous. By using a joint CI, we ensure that the subgroups have robust

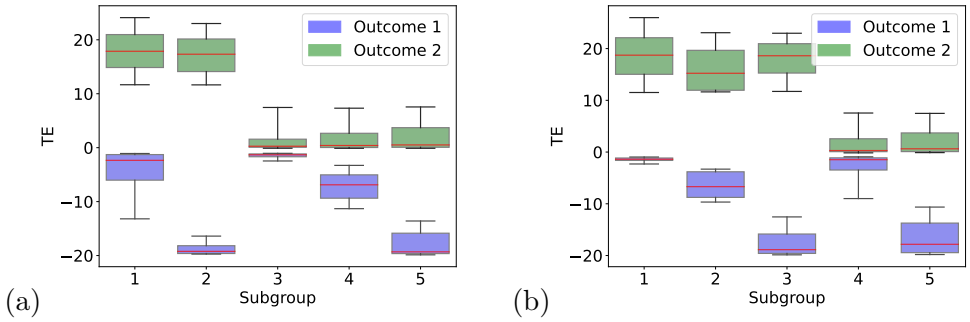

Figure 1: **Jointly partitioned subgroups on synthetic data (MOP-JCI).** Subgroups defined in each outcome when partitioned using the RF estimator on the SCR method (a) and the SCQR method (b). Whiskers show the 25th and 75th percentiles of the treatment effect (TE).

Baselines (R2P)											
		Outcome 1					Outcome 2				
	Num groups	V_{across}	V_{within}	PEHE	CI Width	Cov	V_{across}	V_{within}	PEHE	CI Width	Cov
CMGP on outcome 1	5.50 ±0.19	51.07 ±1.63	**1.64 ±0.11**	0.20 ±0.02	1.22 ±0.13	98.62 ±0.47	3.09 ±1.60	74.53 ±1.70	-	-	-
CMGP on outcome 2	5.03 ±0.18	4.13 ±3.10	52.73 ±2.98	-	-	-	68.97 ±2.07	**4.28 ±1.19**	0.62 ±0.16	3.83 ±1.12	97.97 ±0.59
RF on outcome 1	5.50 ±0.21	**52.51 ±1.65**	1.90 ±0.20	0.58 ±0.04	4.13 ±0.37	98.58 ±0.64	1.96 ±0.48	76.63 ±1.73	-	-	-
RF on outcome 2	4.90 ±0.18	3.77 ±2.23	52.79 ±2.02	-	-	-	**69.49 ±2.44**	7.07 ±1.04	1.08 ±0.10	9.97 ±1.01	98.62 ±0.73

Jointly Partitioned (MOP-JCI)												
				Outcome 1				Outcome 2				
	Num groups	Split Acc	Split Err	V_{across}	V_{within}	PEHE	CI Width	V_{across}	V_{within}	PEHE	CI Width	Cov (joint)
CMGP (SCR)	4.97 ±0.37	97%	13%	46.51 ±3.68	9.67 ±3.67	0.18 ±0.02	1.17 ±0.14	62.71 ±4.81	13.60 ±4.81	1.06 ±0.29	7.34 ±2.26	97.97 ±0.45
RF (SCR)	4.80 ±0.25	100%	17%	48.37 ±1.48	7.58 ±1.15	0.57 ±0.05	4.48 ±0.35	**65.98 ±2.13**	**11.30 ±0.70**	1.10 ±0.10	11.62 ±1.08	98.73 ±0.65
QRF (SCQR)	4.77 ±0.21	100%	10%	**48.92 ±0.94**	**6.11 ±0.37**	0.61 ±0.05	6.08 ±0.49	62.66 ±1.62	11.91 ±0.65	1.18 ±0.10	16.20 ±1.44	98.62 ±0.86

Table 1: Results from synthetic data. We take the mean and standard deviation of each metric across 30 runs. Num groups is the number of subgroups generated. Best performance for V_{across} and V_{within} in each column are in bold.

coverage across the multiple outcomes, regardless if the outcomes show similar or opposing treatment effects.

To evaluate our approach, we tested a baseline method that partitioned the covariate space solely on single outcomes, and demonstrated the shortfall by observing poor heterogeneity across subgroups for both outcomes. Our method showed that we can partition considering the variance across subgroups and within subgroups for each outcome simultaneously, as compared to when they are partitioned on each outcome separately. Additionally, we ensured the validity of our results by reporting the joint coverage and the percentage of correct and incorrect covariate splits. We showed how the tuning parameter used in our method can be

Baselines (R2P)											
		Outcome 1					Outcome 2				
	Num groups	V_{across}	V_{within}	PEHE	CI Width	Cov	V_{across}	V_{within}	PEHE	CI Width	Cov
CMGP on outcome 1	4.17	**18.02**	**21.51**	2.59	14.12	95.38	0.09	0.81	-	-	-
	±0.17	**±2.29**	**±1.51**	±0.19	±1.34	±1.21	±0.04	±0.04			
CMGP on outcome 2	4.10	3.84	39.62	-	-	-	**0.57**	**0.27**	0.27	1.70	97.93
	±0.20	±1.65	±3.24				**±0.03**	**±0.02**	±0.03	±0.20	±0.51
RF on outcome 1	2.47	11.85	29.22	4.06	26.07	97.64	0.01	0.84	-	-	-
	±0.29	±2.63	±2.78	±0.20	±2.47	±0.68	±0.01	±0.04			
RF on outcome 2	1.20	0.61	37.63	-	-	-	0.03	0.82	0.66	4.20	99.38
	±0.15	±1.08	±2.59				±0.04	±0.05	±0.03	±0.20	±0.31

Jointly Partitioned (MOP-JCI)												
				Outcome 1				Outcome 2				
	Num groups	Split Acc	Split Err	V_{across}	V_{within}	PEHE	CI Width	V_{across}	V_{within}	PEHE	CI Width	Cov (joint)
CMGP (SCR)	4.17	43%	87%	15.39	25.10	2.63	12.69	**0.33**	**0.52**	0.25	1.68	93.62
	±0.14			±1.93	±2.51	±0.18	±1.20	**±0.08**	**±0.07**	±0.02	±0.18	±1.03
RF (SCR)	3.83	33%	90%	13.07	25.91	4.01	21.67	0.13	0.73	0.67	3.38	94.93
	±0.26			±1.65	±2.65	±0.20	±2.00	±0.05	±0.06	±0.02	±0.16	±1.00
QRF (SCQR)	4.37	53%	87%	**17.60**	**23.56**	3.95	21.25	0.32	0.55	0.55	3.07	95.64
	±0.23			**±2.04**	**±1.92**	±0.20	±0.97	±0.05	±0.06	±0.02	±0.10	±0.93

Table 2: Results from semi-synthetic data. We take the mean and standard deviation of each metric across 30 runs. Num groups is the number of subgroups generated. Best performance for V_{across} and V_{within} in each column are in bold.

used to favor one outcome over another. Our method paves the way for future work focusing on statistically advanced methods of analyzing RCT data where there are multivariate and multi-output effects. Lastly, we implemented a quantile ITE estimator to partition the data, which reduced the need for a tuning parameter in the algorithm and recalibration at each split, reducing computational burden.

Limitations

Though we used semi-synthetic data with real-world covariates, we only evaluated our data on scenarios where the outcomes are synthetically generated. Future effort should be made towards evaluating the performance of our method on real-world outcomes. In the current approach, we use the Bonferroni correction to adjust the miscoverage rate for both the joint CIs on the ITE estimate for a single outcome, and the joint ITE estimate for multiple outcomes. This method for estimating joint CIs can be conservative, especially as the correlation between outcomes grows. Future work should focus on more precise joint CI estimation. Lastly, future work should investigate how performance changes when there are more than two outcomes.

7. Broader Impacts and Ethics

Our method has the potential to advance precision medicine by identifying subgroups where the response is homogeneous across multiple outcomes. Despite the potential for positive impact as a result of our work, we note a few potential ethics considerations. Primarily, the results of our work are not meant to be interpreted as definitive conclusions drawn about subsets of patients, but rather meant to allow clinicians to propose hypotheses for further investigation. By identifying which covariates determine the subgroups, our method has the potential to serve as a tool to help researchers form testable hypotheses about which patients may be ideal candidates for a therapy.

Acknowledgments

E.H. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064.

References

Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf.

Mohamed Alosh, Kathleen Fritsch, Mohammad Huque, Kooros Mahjoob, Gene Pennello, Mark Rothmann, Estelle Russek-Cohen, Fraser Smith, Stephen Wilson, and Lilly Yue. Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research*, 7:286–303, 10 2015. ISSN 1946-6315. doi: 10.1080/19466315.2015.1077726.

Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*, 355:1064–1069, 3 2000. ISSN 01406736. doi: 10.1016/S0140-6736(00)02039-0.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353–7360, 7 2016. URL https://doi.org/10.1073/pnas.1510489113.

Alfred A. Baumeister and Verne R. Bacharach. A critical analysis of the infant health and development program. *Intelligence*, 23:79–104, 9 1996. ISSN 01602896. doi: 10.1016/S0160-2896(96)90007-0.

John H. Beigel, Kay M. Tomashek, Lori E. Dodd, Aneesh K. Mehta, Barry S. Zingman, Andre C. Kalil, Elizabeth Hohmann, Helen Y. Chu, Annie Luetkemeyer, Susan Kline, Diego Lopez de Castilla, Robert W. Finberg, Kerry Dierberg, Victor Tapson, Lanny Hsieh, Thomas F. Patterson, Roger Paredes, Daniel A. Sweeney, William R. Short, Giota Touloumi, David Chien Lye, Norio Ohmagari, Myoung don Oh, Guillermo M. Ruiz-Palacios, Thomas Benfield, Gerd Fätkenheuer, Mark G. Kortepeter, Robert L. Atmar, C. Buddy Creech, Jens Lundgren, Abdel G. Babiker, Sarah Pett, James D. Neaton, Timothy H. Burgess, Tyler Bonnett, Michelle Green, Mat Makowski, Anu Osinusi, Seema Nayak, and H. Clifford Lane. Remdesivir for the treatment of covid-19 — final report. *New England Journal of Medicine*, 383:1813–1826, 11 2020. ISSN 0028-4793. doi: 10.1056/NEJMoa2007764.

C. S. Berkey, J. J. Anderson, and D. C. Hoaglin. Multiple-outcome meta-analysis of clinical trials. *Statistics in Medicine*, 15:537–557, 3 1996. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19960315)15:5⟨537::AID-SIM176⟩3.0.CO;2-S.

Jaykaran Charan, Rimple Jeet Kaur, Pankaj Bhardwaj, Mainul Haque, Praveen Sharma, Sanjeev Misra, and Brian Godman. Rapid review of suspected adverse drug events due to remdesivir in the who database; findings and implications. *Expert Review of Clinical Pharmacology*, 14:95–103, 1 2021. ISSN 1751-2433. doi: 10.1080/17512433.2021.1856655.

Ruth T. Gross. Infant health and development program (ihdp): Enhancing the outcomes of low birth weight, premature infants in the united states, 1985-1988. *Inter-university Consortium for Political*

and Social Research, 10 1993. URL https://doi.org/10.3886/ICPSR09795.v1.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 1 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.08162.

Hayley E Jones, David I Ohlssen, Beat Neuenschwander, Amy Racine, and Michael Branson. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8:129–143, 4 2011. ISSN 1740-7745. doi: 10.1177/1740774510396933.

Edward H Kennedy, Shreya Kangovi, and Nandita Mitra. Estimating scaled treatment effects with multiple outcomes. *Statistical Methods in Medical Research*, 28: 1094–1104, 4 2019. ISSN 0962-2802. doi: 10.1177/0962280217747130.

Manoj Kumar. Scikit-garden/scikit-garden: A garden for scikit-learn compatible trees, 2017. URL https://github.com/scikit-garden/scikit-garden.

Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, and Mihaela van der Schaar. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *Proceedings of the 34th Conference on Neural Information Processing Systems*, 6 2020.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 7 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1307116.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83:911–938, 2021.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, 10 2011.

Gene Pennello and Mark Rothmann. Bayesian subgroup analysis with hierarchical models. *Biopharmaceutical Applied Statistics Symposium : Volume 2 Biostatistical Analysis of Clinical Trials*, pages 175–192, 2018. doi: 10.1007/978-981-10-7826-2_10. URL https://doi.org/10.1007/978-981-10-7826-2_10.

Alexandros Rekkas, Jessica K. Paulus, Gowri Raman, John B. Wong, Ewout W. Steyerberg, Peter R. Rijnbeek, David M. Kent, and David van Klaveren. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Medical Research Methodology*, 20:264, 12 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-01145-1.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, page 11, 2019.

Heidi Seibold, Achim Zeileis, and Torsten Hothorn. Model-Based recursive partitioning for subgroup analyses. *The international journal of biostatistics*, 12(1):45–63, 5 2016.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158, 2009.

Ravi Varadhan, Jodi B. Segal, Cynthia M. Boyd, Albert W. Wu, and Carlos O. Weiss. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology*, 66:818–825, 8 2013. ISSN 08954356. doi: 10.1016/j.jclinepi.2013.02.009.

Victoria Vickerstaff, Gareth Ambler, Michael King, Irwin Nazareth, and Rumana Z Omar. Are multiple primary outcomes analysed appropriately in randomised controlled trials? a review. *Contemporary Clinical Trials*, 45:8–12, 2015.

Han Wu, Sarah Tan, Weiwei Li, Mia Garrard, Adam Obeng, Drew Dimmery, Shaun Singh, Hanson Wang, Daniel Jiang, and Eytan Bakshy. Interpretable personalized experimentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pages 4173–4183, New York, NY, USA, August 2022. Association for Computing Machinery.

Leon Yao, Caroline Lo, Israel Nir, Sarah Tan, Ariel Evnine, Adam Lerer, and Alex Peysakhovich. Efficient heterogeneous treatment effect estimation with multiple experiments and multiple outcomes. June 2022.

Frank B. Yoon, Garrett M. Fitzmaurice, Stuart R. Lipsitz, Nicholas J. Horton, Nan M. Laird, and Sharon-Lise T. Normand. Alternative methods for testing treatment effects on the basis of multiple outcomes: Simulation and case study. *Statistics in Medicine*, 30:1917–1932, 7 2011. ISSN 02776715. doi: 10.1002/sim.4262.

Appendix A. Datasets

A.1. Synthetic Data

This section describes the synthetic data that we adapted from Lee et al. (2020). We use a logistic function to generate both outcome distributions. The outputs were generated using the following equations to simulate for the control and treated populations, in the first outcome:

$$\text{Control: } y_A(0) \sim \mathcal{N}(X_0\beta + (1 + e^{-(x_j - m_j)})^{-1} + 20, 0.1)$$
$$\text{Treatment: } y_A(1) \sim \mathcal{N}(X_0\beta + 20(1 + e^{-(x_j - m_j)})^{-1}, 0.1)$$

and the second:

$$\text{Control: } y_B(0) \sim \mathcal{N}(X_0\beta + (1 + e^{-(x_j - m_j)})^{-1} + x_j, 0.1)$$
$$\text{Treatment: } y_B(1) \sim \mathcal{N}(X_0\beta + x_j(1 + e^{-(x_j - m_j)})^{-1} + x_j, 0.1)$$

where m_j applies a shift by the mean of the covariate x_j, X_0 represents a matrix of all the covariate values except for the covariate x_j and β applies coefficients that are randomly sampled among $(0, 0.1, 0.2, 0.3, 0.4)$ probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$ respectively. Table S1 shows the distributions of the simulated covariates.

This dataset was modified to assess for the robustness of our model with two additional variations of this dataset. One showed the effect of correlated covariates, specifically time and ALT levels. This accordingly results in correlated outcomes. The other showed a heteroscedastic trend for time on one outcome and ALT levels on the other. In all synthetic datasets, we created a set of 300 and 200 samples for training and testing respectively.

The treatment effect trends for the synthetic data across each covariate are shown in Figure S1 (Figure S2 for correlated covariates, and Figure S3 for heteroscedastic data). The x-axis shows the values across each respective covariate and the y-axis shows the value of each respective outcome. Note the correlation between time and ALT level in each outcome.

A.2. Semi-synthetic Data

The first outcome showed the cognitive development score assessed by the Stanford-Binet Intelligence Scale, where infants enrolled in the intervention showed higher mean scores than the infants in the control population. Score differences were found to be dependent on the nnhealth of the infant. The second outcome evaluated the health status score of the infant measured by the mothers' report on the morbidity index. The health score showed positive treatment effect, but was dependent on the mother's age where younger mothers tended to report more frequent adverse health conditions. As part of the Response Surface B, other covariates in the dataset are randomly assign to have a small effect of the outcomes.

Treatment effects for the semi-synthetic data across each covariate are shown in Figure S4. The x-axis details the values across each respective covariate and the y-axis shows the value of each respective outcome. Note the relationship between nnhealth and Outcome 1, and momage and Outcome 2. The dataset consisted of 608 untreated and 139 treated subjects, where the training and testing sets were split by 80% and 20%, respectively. The dataset included 25 covariates.

age	$\sim \mathcal{N}(66, 4.1)$
white blood cell count (x10^9 per L)	$\sim \mathcal{N}(6.2, 1)$
lymphocyte count (x10^9 per L)	$\sim \mathcal{N}(0.8, 0.1)$
platelet count (x10^9 per L)	$\sim \mathcal{N}(183, 20.4)$
serum creatinine (U/L)	$\sim \mathcal{N}(68, 6.6)$
asparatate aminotransferase (U/L)	$\sim \mathcal{N}(31, 5.1)$
alanine aminotransferase (U/L)	$\sim \mathcal{N}(16, 5.1)$
lactate dehydrogenase (U/L)	$\sim \mathcal{N}(339, 51)$
creatine kinase (U/L)	$\sim \mathcal{N}(76, 21)$
time from symptom onset to starting the trial (days)	$\sim Unif(9, 14)$

Table S1: Distributions of covariates in synthetic data.

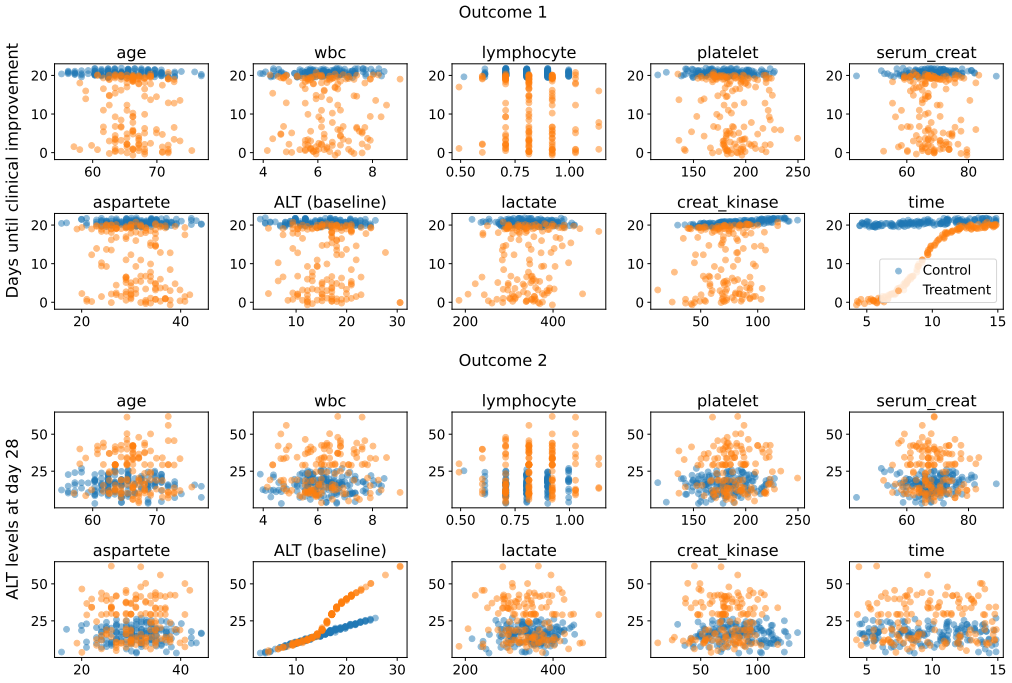

Figure S1: **Treatment effect across covariates in the synthetic dataset.** Distributions are split across control and treatment populations, shown in blue and orange respectively.

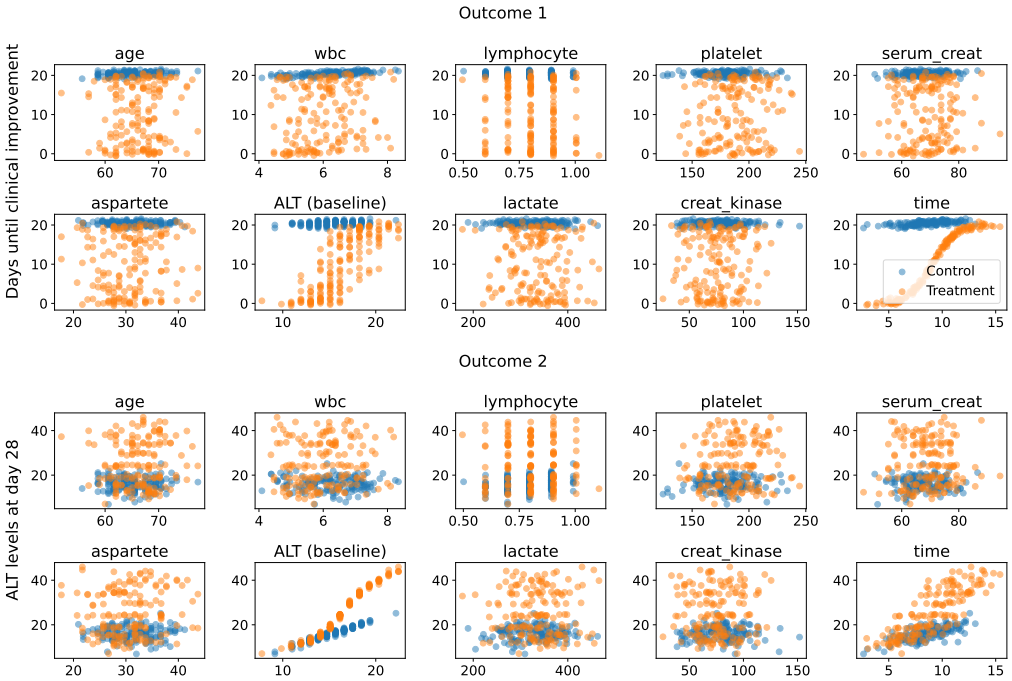

Figure S2: **Treatment effect across covariates in the synthetic dataset, when there are correlated covariates.** Distributions are split across control and treatment populations, shown in blue and orange respectively.

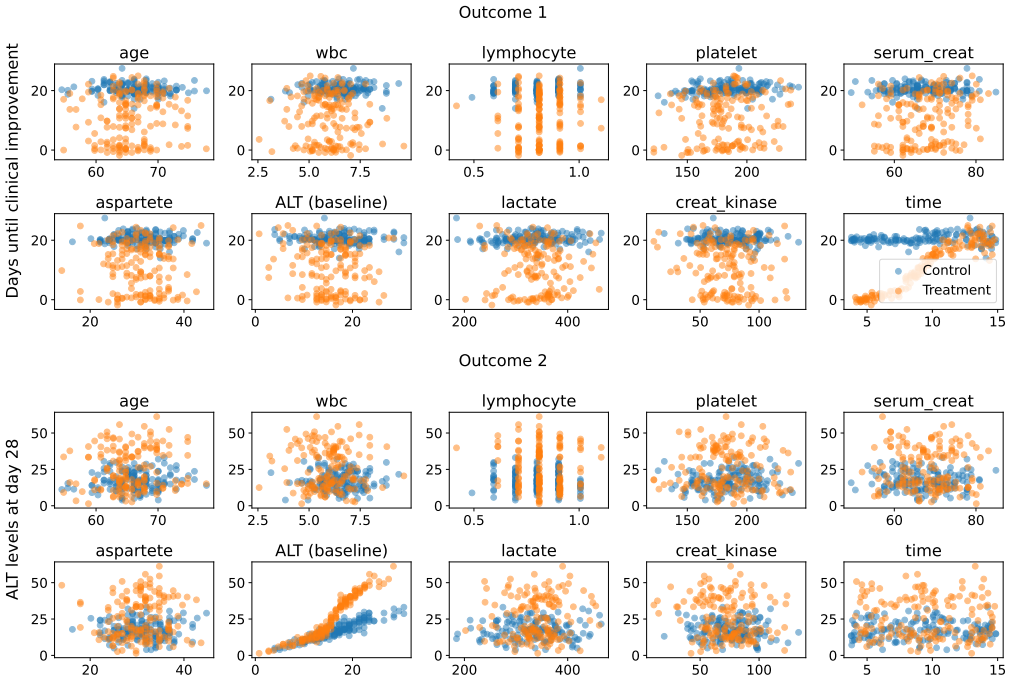

Figure S3: **Treatment effect across covariates in the synthetic dataset, where the data is heteroscedastic.** Distributions are split across control and treatment populations, shown in blue and orange respectively.

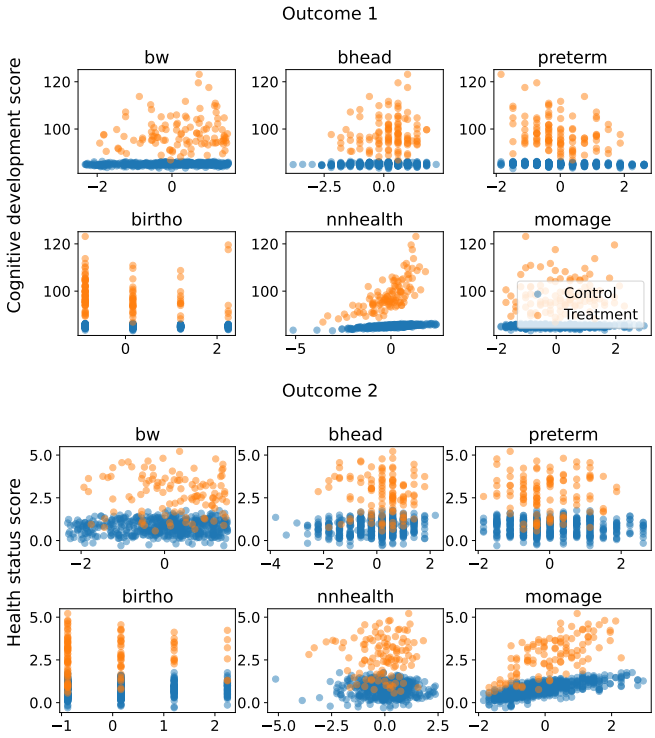

Figure S4: **Treatment effect across covariates in the semi-synthetic dataset.** Distributions are split across control and treatment populations, shown in blue and orange respectively. Distributions are only shown in the 6 continuous covariates (remaining covariates are binary). Note. the continuous covariate values are normalized.

Appendix B. ITE Estimators

The ITE estimators, Random Forest (RF) and Quantile Random Forest (QRF), were adapted from their original implementation in scikit-learn (Pedregosa et al., 2011) and scikit-garden (Kumar, 2017) respectively. The adaptations were made in order to work in a conformal prediction framework. The results of the hyperparameter tuning for the RF on the synthetic data are found in Table S2; the hyperparameters for the semi-synthetic data are found in Table S3. We chose to not tune the QRF, as the default parameters already gave high precision estimates.

n_estimators	450
random_state	0
min_samples_split	2
min_samples_leaf	1
max_depth	38
max_features	auto
bootstrap	True

Table S2: Hyperparameters used in RF ITE estimator for the synthetic data.

n_estimators	450
random_state	None
min_samples_split	3
min_samples_leaf	1
max_depth	50
max_features	sqrt
bootstrap	False

Table S3: Hyperparameters used in RF ITE estimator for the semi-synthetic data.

Appendix C. Methodology

C.1. Assumptions

The assumptions on which our methodology is based are as follows:

1. This methodology holds in a RCT environment, where all patients have the same set of covariates and outcomes and there is no missingness.

2. The outcomes are continuous values.

3. There is an expected heterogeneous behavior in the dataset (for example in the medical space, the behavior may be proved through clinical references).

4. Our methodology can be tested on multiple outcomes ≥ 2, though for the purposes of this study, we focus on 2 outcomes.

C.2. Extension to Multiple Outcomes

We provide an alternative formulation of the objective function below that can work in the setting of d outcomes, where τ_i is the weight for outcome i. The equations for the objective functions for SCR and SCQR are below.

$$\text{SCR: minimize} \sum_{g \in \Pi} \lambda \left(\sum_{i \in d} (\tau_i W_g^i) \right) + (1 - \lambda) \left(\sum_{i \in d} (\tau_i V_g^i) \right)$$

$$\text{SCQR: minimize} \sum_{g \in \Pi} \sum_{i \in d} (\tau_i V_g^i)$$

C.3. Partitioning Algorithm for SCR

The partitioning algorithm for SCR can be found in Algorithm 2.

Algorithm 2: SCR Recursive Partitioning on Two Outcomes

Input: G_{node}, data
for *Covariate j in data* **do**
  for *Unique value x of covariate j* **do**
    Split data in two branches on x
    Compute V^A and W^A for each branch
    Compute V^B and W^B for each branch
    Set G_{split} to G_{node} less the sum of $\lambda(\beta W^A + (1-\beta)W^B) + (1-\lambda)(\beta V^A + (1-\beta)V^B)$ from each branch
  end
end
Save best G_{split} with covariate j
if $G_{split} > \gamma G_{node}$ **then**
  $G_{node} \leftarrow G_{split}$
  Partition on each branch
else
  Set current node to leaf
end

C.4. Hyperparameters

Hyperparameter tuning was conducted on the ITE estimators using random forests as shown in Appendix B. As for the hyperparameters in the partitioning algorithm, λ and γ were tested across varying values and set to the values shown in Table S4 (such that overall, V_across was maximized, V_in was minimized and ci_width was minimized). λ is used to vary the weight between the expected absolute deviation within a group and the CI width, affecting the number of subgroups and the inter- and intra-subgroup variance. γ controls for regularization, where too small of a value can lead to overfitting with a large number of subgroups and too large of a value can lead to poor performance with a small number of subgroups. The effects of varying λ,and γ in the synthetic dataset is shown

in Figure S5 (tuning for the semi-synthetic dataset is shown in Figure S6).

In our experiments, β determines the weight of each outcome in the algorithm. In the two outcome case that we have explored, a β value other than .5 will favor one outcome over the other. This parameter can be used to prioritize partitioning on one outcome more than the other. Additionally, in scenarios where the magnitudes of the treatment effects of each outcome are very different, β can be used to weight the effect accordingly. We tested the impact that β has on certain metrics for both datasets. Figure S7 shows the effect of β on the performance metrics for the synthetic dataset. Figure S8 shows the effect in the IHDP dataset. Since in the IHDP dataset, outcome 1 has a higher magnitude than outcome 2, setting β to be a lower than .5 allows the algorithm to find heterogeneity in outcome 2. Since many covariates contribute to both outcomes in IHDP, the error metric is not reported.

To generate the confidence regions W and V, we use the miscoverage rates of .1 and .8, respectively.

	Synthetic	Semi-synthetic
λ	0.25	0
γ	0.05	0.02
β	0.5	0.25

Table S4: Hyperparameters used in the partitioning algorithm for each dataset.

C.5. Licenses

The license of the assets used in this paper are as follows:

- Robust recursive partitioning algorithm, and synthetic data:https://github.com/vanderschaarlab/mlforhealthlabpub/blob/main/LICENSE.md
- SCR:https://github.com/ryantibs/conformal/blob/master/LICENSE
- SCQR:https://github.com/yromano/cqr/blob/master/LICENSE
- IHDP dataset: A license was not provided, the code was zipped in the supplementary material of (Hill, 2011).

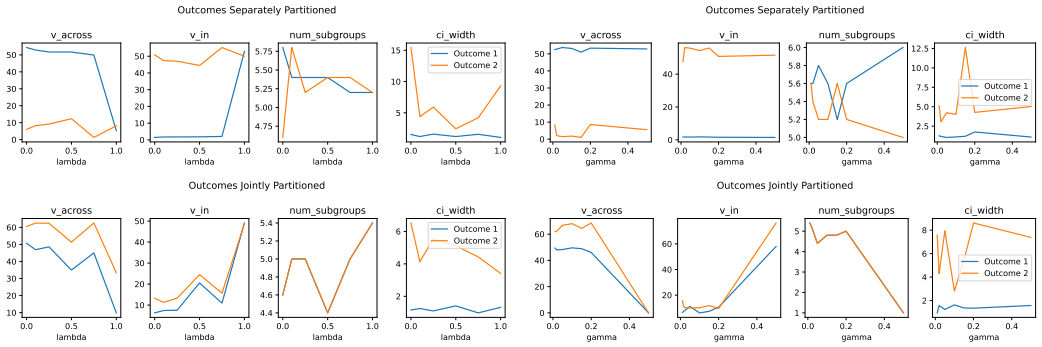

Figure S5: Performance across separately and jointly partitioned outcomes varying λ and γ hyperparameters in the synthetic dataset.

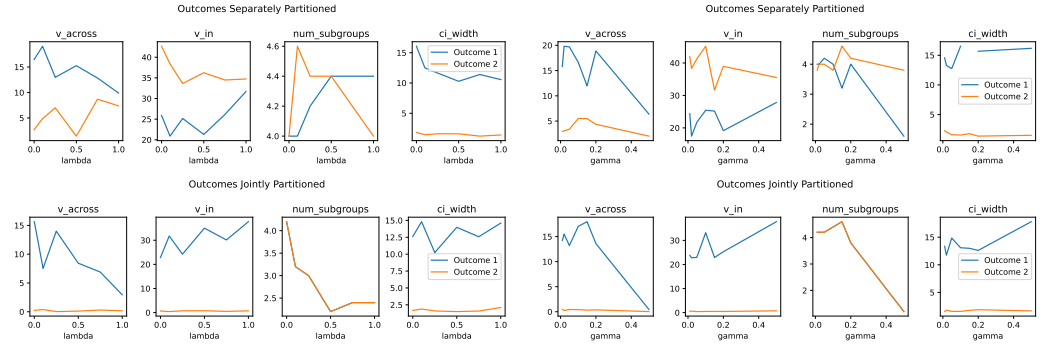

Figure S6: Performance across separately and jointly partitioned outcomes varying λ and γ and hyperparameters in the semi-synthetic dataset.

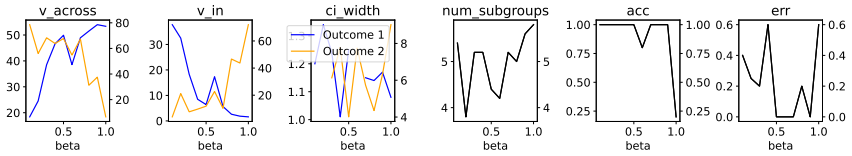

Figure S7: β tuning using SCR CMGP on the synthetic data

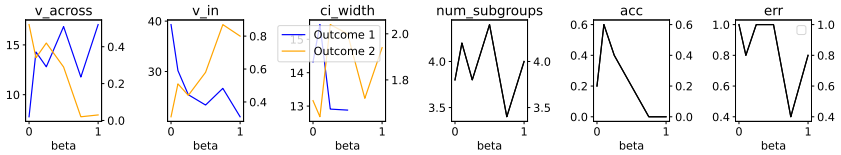

Figure S8: β tuning using SCR CMGP on the semi-synthetic data

Appendix D. Additional Results

In this section we show additional results that were omitted from the paper. We show the tables and figures for the results on the two versions from the synthetic data that were not discussed in the paper. These are correlated covariates and heteroscedastic data. We additionally show subgroup analyses for all results on the synthetic and semi-synthetic data. In these subgroup analyses we show a sample subgroup from each method and the characteristics of each subgroup.

D.1. Separate Versus Joint Partitioning

Separately partitioned subgroups using the CMGP estimator (Baseline R2P) are shown on synthetic data (Figure S9) and on semi-synthetic data (Figure S10). Jointly partitioned subgroups on semi-synthetic data using our method (MOP-JCI) are shown in Figure S11.

D.2. Correlated Covariates

Here, we show the results from the synthetic data with correlated covariates. Figure S12 shows the treatment effect for both outcomes of each subgroup when partitioned on each outcome separately. We show the CMGP ITE estimator method. Figure S13 shows the treatment effect for both outcomes of each subgroup when partitioned on each outcome jointly. We show the method using SCR and SCQR, both using a RF estimator.

In Table S5, we show the numerical results of each subgroup with mean and standard deviation when running each method 30 times. We show the variance across each group, the variance within each group, the precision of the ITE estimator, the coverage, and the average CI of each subgroup. The PEHE is computed using the 50th quantile for the SCQR method.

D.3. Heteroscedasticity

Here, we show the results from the synthetic data with added heteroscedasticity. Figure S14 shows the treatment effect for both outcomes of each subgroup when partitioned on each outcome separately. We show the CMGP ITE estimator method. Figure S15 shows the treatment effect for both outcomes of each subgroup when partitioned on each outcome jointly. We show the method using SCR and SCQR, both using a RF estimator.

In Table S6 we show the numerical results of each partitioning with mean and standard deviation when running each method 30 times. We show the variance across each group, the variance within each group, the precision of the ITE estimator, the coverage, and the average CI of each subgroup. The PEHE is computed using the 50th quantile for the SCQR method.

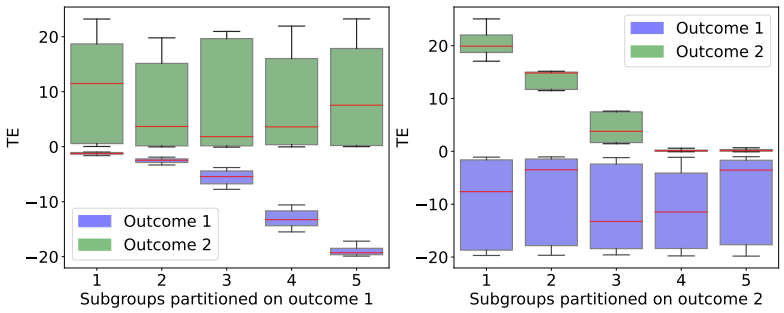

Figure S9: **Separately partitioned subgroups on synthetic data using the CMGP estimator (Baseline R2P).** Subgroups defined in each outcome when partitioned on a single treatment outcome individually. Left plot shows the box plot for treatment effects when the covariates are partitioned using outcome 1 only. The right plot shows the box plots when the covariates are partitioned using outcome 2 only. Whiskers show the 25th and 75th percentiles.

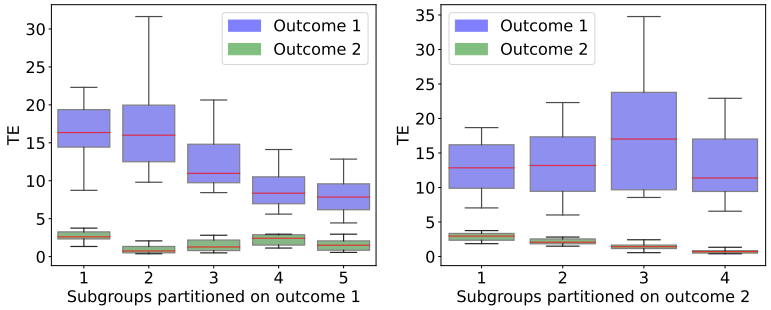

Figure S10: **Separately partitioned subgroups on semi-synthetic data using the CMGP estimator (Baseline R2P).** Subgroups defined in each outcome when partitioned on a single treatment outcome individually. Left plot shows the box plot for treatment effects when the covariates are partitioned using outcome 1 only. The right plot shows the box plots when the covariates are partitioned using outcome 2 only. Whiskers show the 25th and 75th percentiles.

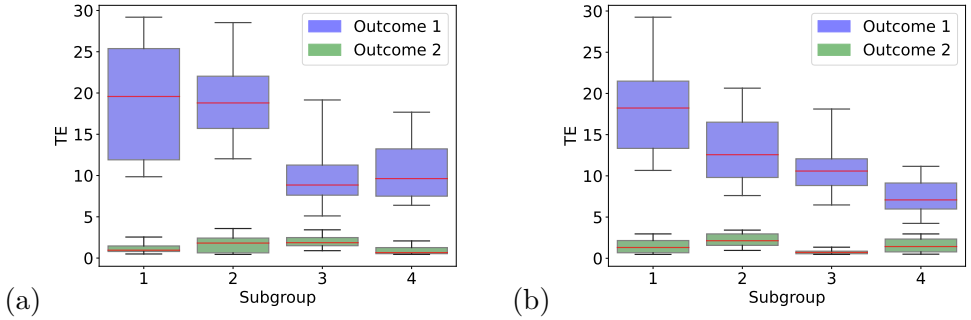

Figure S11: **Jointly partitioned subgroups on semi-synthetic data (MOP-JCI).** Subgroups defined in each outcome when partitioned using the RF estimator on the SCR method (a) and the quantile method (b). Whiskers show the 25th and 75th percentiles.

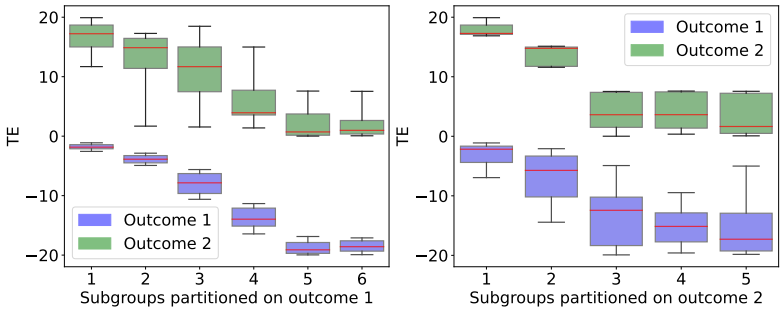

Figure S12: **Separately partitioned subgroups on synthetic data with correlated covariates using the CMGP estimator (Baseline R2P).** Subgroups defined in each outcome when partitioned on a single treatment outcome individually. Left plot shows the box plot for treatment effects when the covariates are partitioned using outcome 1 only. The right plot shows the box plots when the covariates are partitioned using outcome 2 only. Whiskers show the 25th and 75th percentiles.

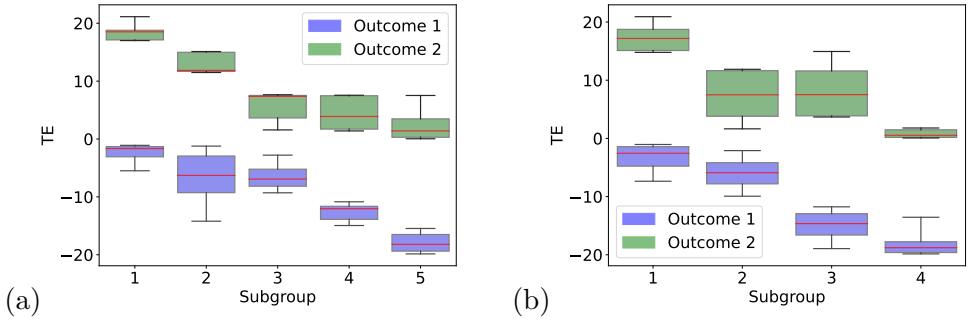

Figure S13: **Jointly partitioned subgroups on synthetic data with correlated covariates (MOP-JCI).** Subgroups defined in each outcome when partitioned using the RF estimator on the SCR method (left) and the quantile method (right). Whiskers show the 25th and 75th percentiles.

Baselines (R2P)											
		Outcome 1					Outcome 2				
	Num groups	V_{across}	V_{within}	PEHE	CI Width	Cov	V_{across}	V_{within}	PEHE	CI Width	Cov
CMGP on outcome 1	5.17	**40.68**	**1.57**	0.41	1.94	97.95	23.92	18.09	-	-	-
	±0.24	**±1.11**	**±0.13**	±0.11	±0.28	±0.59	±1.14	±0.83			
CMGP on outcome 2	5.03	25.90	16.44	-	-	-	**40.21**	**2.43**	0.35	2.56	98.62
	±0.18	±1.23	±1.04				**±1.81**	**±0.59**	±0.09	±0.64	±0.41
RF on outcome 1	5.03	40.11	1.70	0.51	3.77	99.45	24.36	18.11	-	-	-
	±0.25	±1.51	±0.12	±0.03	±0.34	±0.35	±1.58	±0.77			
RF on outcome 2	5.17	24.54	17.21	-	-	-	37.79	4.02	0.66	6.45	99.08
	±0.24	±1.81	±1.24				±1.74	±0.76	±0.07	±0.56	±0.55

Jointly Partitioned (MOP-JCI)												
				Outcome 1				Outcome 2				
	Num groups	Split Acc	Split Err	V_{across}	V_{within}	PEHE	CI Width	V_{across}	V_{within}	PEHE	CI Width	Cov (joint)
CMGP (SCR)	5.00	97%	17%	35.88	5.35	0.48	1.80	**33.93**	**7.91**	0.39	2.67	97.13
	±0.22			±1.58	±0.76	±0.19	±0.31	**±1.93**	**±1.29**	±0.08	±0.80	±0.74
RF (SCR)	5.03	97%	27%	**35.94**	4.56	0.52	3.90	31.89	9.52	0.67	6.60	99.55
	±0.25			**±1.48**	±0.73	±0.04	±0.40	±2.01	±1.42	±0.09	±0.68	±0.28
QRF (SCQR)	4.97	93%	13%	35.21	**4.01**	0.60	5.63	30.62	9.95	0.72	9.27	97.55
	±0.18			±1.70	**±0.60**	±0.04	±0.41	±1.94	±1.63	±0.07	±0.82	±1.65

Table S5: Results from synthetic data with correlated covariates. We take the mean and standard deviation of each metric across 30 runs. Num groups is the number of subgroups generated. Best performance for V_{across} and V_{within} in each column are in bold.

Baselines (R2P)											
		Outcome 1					Outcome 2				
	Num groups	V_{across}	V_{within}	PEHE	CI Width	Cov	V_{across}	V_{within}	PEHE	CI Width	Cov
CMGP on outcome 1	5.20	45.71	15.27	2.48	14.74	96.80	8.03	84.02	-	-	-
	±0.25	±2.80	±2.77	±0.11	±0.74	±0.94	±6.20	±6.21			
CMGP on outcome 2	5.23	13.28	48.92	-	-	-	38.02	51.93	4.02	22.98	97.75
	±0.19	±5.54	±5.85				±8.75	±9.10	±0.23	±1.40	±0.48
RF on outcome 1	4.90	**48.56**	**12.85**	2.51	14.49	96.67	10.54	80.38	-	-	-
	±0.27	**±2.57**	**±2.10**	±0.07	±0.68	±0.92	±5.06	±5.96			
RF on outcome 2	5.20	12.32	49.89	-	-	-	**41.08**	**50.11**	4.30	24.79	97.45
	±0.21	±4.82	±4.96				**±7.71**	**±7.44**	±0.20	±1.45	±0.88

Jointly Partitioned (MOP-JCI)												
				Outcome 1				Outcome 2				
	Num groups	Split Acc	Split Err	V_{across}	V_{within}	PEHE	CI Width	V_{across}	V_{within}	PEHE	CI Width	Cov (joint)
CMGP (SCR)	5.20	100%	53%	46.44	14.91	2.42	13.69	59.23	33.33	4.19	23.94	97.17
	±0.21			±2.26	±2.33	±0.08	±0.43	±5.95	±5.19	±0.30	±1.58	±0.63
RF (SCR)	5.17	100%	70%	47.68	12.96	2.45	13.51	53.57	38.03	4.42	25.88	97.13
	±0.22			±1.80	±1.16	±0.07	±0.69	±4.57	±4.27	±0.18	±1.27	±0.66
QRF (SCQR)	4.90	100%	20%	**48.88**	**10.92**	2.62	15.54	**62.58**	**26.43**	4.40	28.51	97.13
	±0.23			**±1.13**	**±0.64**	±0.09	±1.03	**±3.86**	**±2.86**	±0.16	±1.56	±0.98

Table S6: Results from heteroscedastic synthetic data. We take the mean and standard deviation of each metric across 30 runs. Num groups is the number of subgroups generated. Best performance for V_{across} and V_{within} in each column are in bold.

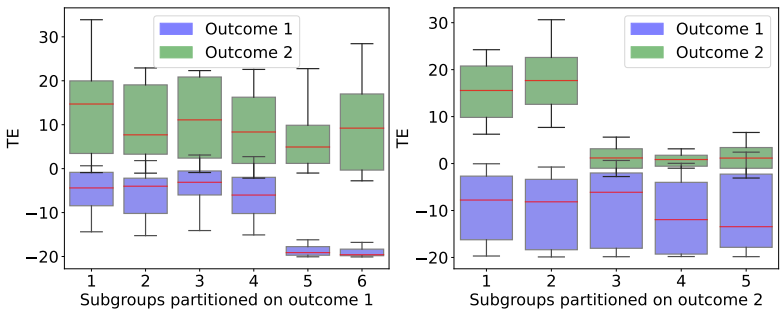

Figure S14: **Separately partitioned subgroups on heteroscedastic synthetic data using the CMGP estimator (Baseline R2P).** Subgroups defined in each outcome when partitioned on a single treatment outcome individually. Left plot shows the box plot for treatment effects when the covariates are partitioned using outcome 1 only. The right plot shows the box plots when the covariates are partitioned using outcome 2 only. Whiskers show the 25th and 75th percentiles.

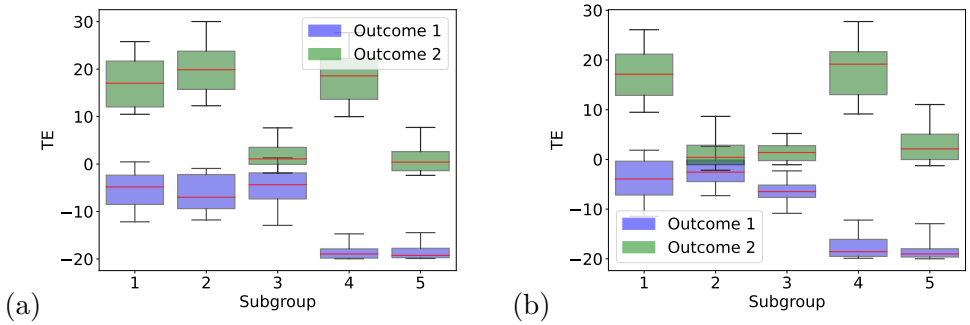

Figure S15: **Jointly partitioned subgroups on heteroscedastic synthetic data (MOP-JCI).** Subgroups defined in each outcome when partitioned using the RF estimator on the SCR method (left) and the SCQR method (right). Whiskers show the 25th and 75th percentiles.

Appendix E. Subgroup Characteristics

In this section, the characteristics of the subgroups formed by each partioning method on each dataset are shown. For the synthetic datasets, we show the statistics of the age and time variable since those covariates determine the outcome distribution. For the semi-synthetic dataset (IHDP), we show the statistics of five of the continuous covariates in the dataset the birthweight (bw), neonatal health (nnhealth), weeks born preterm (preterm), age of mother (momage), and birth head size (bhead).

E.0.1. Synthetic data

Table S7 shows an example of the subgroup characteristics for the separate partitioning methods. Table S8 shows an example of the subgroup characteristics for the joint partitioning methods. Note the uncorrelated outcomes dataset was used in the paper as the synthetic dataset.

E.0.2. Correlated Covariates

Table S9 shows the subgroup characteristics for separately partitioned, and Table S10 shows the subgroup characteristics when jointly partitioned.

E.0.3. Heteroscedasticity

Table S11 shows an example of the subgroup characteristics for the separate partitioning methods, and Table S12 shows an example of the subgroup characteristics for the joint partitioning methods.

E.0.4. Semi-synthetic data

Table S13 shows an example of the subgroup characteristics for the separate partitioning methods. Table S14 shows an example of the subgroup characteristics for the joint partitioning methods.

	SCR CMGP separate on outcome 1						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	49.0	16.03	5.51	13.66	0.83	-1.25	0.24
2.0	27.0	15.18	4.90	11.45	0.36	-2.56	0.50
3.0	29.0	15.31	5.70	10.16	0.47	-5.60	1.63
4.0	28.0	15.25	5.40	8.33	0.43	-13.10	1.79
5.0	67.0	16.34	5.42	5.71	0.95	-18.92	0.93

	SCR CMGP separate on outcome 2						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	60.0	21.99	2.48	9.41	3.28	20.50	2.63
2.0	31.0	17.72	0.52	10.03	3.44	13.60	1.66
3.0	38.0	15.09	0.88	9.13	3.14	4.39	2.58
4.0	34.0	10.01	3.14	8.84	2.92	0.16	0.22
5.0	37.0	10.22	2.94	9.89	3.33	0.21	0.25

	SCR RF separate on outcome 1						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	46.0	16.26	6.00	12.95	1.19	-1.61	0.61
2.0	31.0	16.76	6.40	12.65	1.19	-1.85	0.74
3.0	23.0	16.26	4.98	10.00	0.44	-6.24	1.62
4.0	45.0	15.33	5.20	7.85	0.66	-15.05	2.37
5.0	25.0	15.87	5.63	5.64	0.87	-19.10	0.63
6.0	30.0	14.77	5.82	5.09	0.84	-19.45	0.49

	SCR RF separate on outcome 2						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	46.0	23.55	2.97	9.72	3.31	21.93	2.95
2.0	27.0	17.65	1.26	8.89	2.67	12.72	3.83
3.0	29.0	17.65	1.12	9.12	3.42	12.82	3.51
4.0	72.0	11.38	3.08	10.79	2.59	0.93	1.35
5.0	26.0	10.78	3.16	5.25	0.88	0.62	0.98

Table S7: Subgroup characteristics from separate partitioning on synthetic data (Baseline R2P).

	SCR CMGP joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 0 mean	tau 0std	tau 1 mean	tau 1 std
0	60.0	19.77	3.18	12.08	1.61	-2.98	2.26	16.62	5.60
1	47.0	19.94	2.92	6.89	1.67	-16.41	3.56	17.32	4.97
2	54.0	11.45	3.10	12.19	1.74	-3.15	2.71	0.88	1.30
3	39.0	11.71	2.70	6.26	1.37	-18.10	2.17	0.92	1.35

	SCR RF joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 0 mean	tau 0std	tau 1 mean	tau 1 std
0	72.0	20.10	3.20	11.94	2.03	-4.42	4.25	17.83	4.50
1	28.0	19.77	2.69	6.31	1.11	-18.67	1.27	17.51	4.13
2	32.0	11.93	2.60	13.61	1.04	-1.48	0.49	1.68	2.71
3	27.0	10.74	4.33	10.22	0.75	-7.12	2.85	1.63	2.46
4	41.0	11.65	3.45	6.49	1.49	-17.93	2.37	2.10	3.02

	SCQR RF joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 0 mean	tau 0std	tau 1 mean	tau 1 std
0	39.0	21.19	3.46	13.19	0.91	-1.47	0.45	18.58	4.72
1	17.0	20.09	3.50	10.05	0.75	-6.44	2.60	16.92	4.86
2	45.0	20.78	2.59	6.44	1.45	-17.50	2.70	18.40	3.69
3	51.0	11.83	3.40	12.36	1.72	-2.79	2.56	1.74	2.76
4	48.0	11.65	4.45	6.81	1.63	-16.50	3.40	1.85	2.52

Table S8: Subgroup characteristics from joint partitioning on synthetic data (MOP-JCI).

Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
	SCR CMGP separate on outcome 1						
1.0	33.0	19.17	1.47	12.48	0.84	-1.79	0.49
2.0	30.0	17.45	1.74	10.80	0.32	-3.88	0.77
3.0	42.0	17.12	1.79	9.65	0.43	-7.86	1.82
4.0	46.0	15.64	1.45	8.29	0.43	-13.74	1.74
5.0	30.0	13.19	2.19	5.45	1.72	-18.79	1.08
6.0	19.0	13.55	1.70	6.00	1.41	-18.53	1.04
	SCR CMGP separate on outcome 2						
1.0	41.0	19.69	0.83	11.69	1.31	17.98	1.14
2.0	49.0	17.62	0.50	10.14	1.35	13.38	1.64
3.0	40.0	14.47	1.76	7.88	2.18	3.62	2.76
4.0	27.0	14.57	1.60	7.64	1.67	3.76	2.96
5.0	43.0	14.05	1.90	7.03	2.23	3.12	2.98
	SCR RF separate on outcome 1						
1.0	26.0	19.96	1.10	12.18	1.25	-2.13	1.04
2.0	25.0	16.91	1.00	11.24	0.92	-3.04	1.28
3.0	46.0	16.38	1.67	9.52	0.33	-7.75	1.45
4.0	40.0	15.01	1.56	8.17	0.44	-13.91	1.88
5.0	33.0	14.35	1.57	6.88	0.38	-17.84	0.76
6.0	30.0	12.79	1.67	5.17	1.15	-19.47	0.34
	SCR RF separate on outcome 2						
1.0	32.0	19.90	1.04	11.72	1.50	18.62	1.39
2.0	47.0	17.27	0.49	9.83	1.57	13.02	1.63
3.0	57.0	15.38	0.49	8.53	1.50	5.54	1.86
4.0	64.0	12.94	1.02	6.59	1.63	0.83	0.60

Table S9: Subgroup characteristics from separate partitioning on correlated covariate synthetic data (Baseline R2P).

Subgroup	count	ALT mean	ALT std	time mean	time std	tau 0 mean	tau 0std	tau 1 mean	tau 1 std
	SCR CMGP joint								
0	32.0	19.58	1.13	11.87	1.42	-2.77	1.95	18.22	1.50
1	54.0	17.42	0.47	10.44	1.25	-5.49	3.23	13.38	1.67
2	24.0	15.41	0.82	9.82	0.97	-7.18	3.22	5.54	2.50
3	42.0	16.07	0.86	7.56	0.90	-15.42	2.29	7.87	3.57
4	48.0	13.09	1.18	5.85	1.74	-17.99	2.21	0.83	0.66
	SCR RF joint								
0	33.0	20.31	1.26	12.36	1.52	-2.35	1.46	18.44	1.42
1	51.0	17.53	0.54	10.34	1.72	-6.65	4.33	13.05	1.65
2	29.0	15.37	0.89	10.06	0.72	-6.38	2.16	5.59	2.42
3	21.0	15.02	0.96	8.50	0.33	-12.65	1.41	4.60	2.69
4	66.0	13.24	1.89	6.27	1.45	-17.92	1.61	2.12	2.55
	SCQR RF joint								
0	36.0	19.18	1.50	11.74	1.84	-3.32	2.39	17.37	2.28
1	58.0	15.80	0.95	10.15	0.85	-5.94	2.57	7.44	3.59
2	54.0	15.99	1.00	7.82	0.81	-14.94	2.28	7.94	3.90
3	52.0	12.80	1.07	6.05	1.51	-18.10	2.03	0.76	0.67

Table S10: Subgroup characteristics from joint partitioning on correlated covariate synthetic data (MOP-JCI).

	SCR CMGP separate on outcome 1						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	37.0	18.04	5.31	11.51	1.95	-5.06	4.76
2.0	23.0	16.46	4.42	11.59	2.41	-5.74	5.92
3.0	20.0	16.71	4.41	11.57	2.19	-4.21	5.41
4.0	55.0	16.30	4.69	11.39	2.29	-5.98	5.68
5.0	37.0	14.57	4.81	5.98	1.17	-18.57	1.53
6.0	28.0	17.25	6.45	5.85	1.05	-18.91	1.46

	SCR CMGP separate on outcome 2						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	64.0	19.36	3.13	9.88	2.88	15.76	6.74
2.0	48.0	20.85	3.71	9.71	3.26	18.16	7.75
3.0	31.0	12.58	2.79	10.05	3.59	1.20	3.10
4.0	16.0	11.74	3.32	8.84	3.35	0.73	2.08
5.0	41.0	11.74	2.90	9.37	3.44	1.55	3.20

	SCR RF separate on outcome 1						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	50.0	16.65	4.94	11.85	1.52	-2.94	3.50
2.0	60.0	17.53	4.66	12.45	1.72	-2.52	3.52
3.0	43.0	16.43	5.26	6.77	1.54	-17.09	3.60
4.0	47.0	16.15	5.04	6.97	1.66	-16.43	3.84

	SCR RF separate on outcome 2						
Subgroup	count	ALT mean	ALT std	time mean	time std	tau 1 mean	tau 1 std
1.0	42.0	22.54	2.28	11.63	1.96	22.12	4.94
2.0	14.0	23.64	2.35	5.83	1.20	23.67	5.81
3.0	25.0	16.57	1.94	9.46	2.69	9.77	7.00
4.0	65.0	16.73	1.59	9.90	3.18	9.81	6.67
5.0	30.0	10.55	2.12	9.63	3.58	0.10	1.57
6.0	24.0	10.58	1.56	9.14	2.74	-0.26	1.26

Table S11: Subgroup characteristics from separate partitioning on heteroscedastic synthetic data (Baseline R2P).

	SCR CMGP joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau0 mean	tau 0 std	tau 1 mean	tau1 std
0	30.0	18.66	4.61	12.96	1.37	-2.00	4.24	11.90	11.47
1	25.0	19.01	4.08	12.66	1.35	-3.05	3.91	13.50	9.59
2	19.0	17.85	3.44	9.28	0.65	-8.75	2.84	11.75	9.59
3	36.0	10.12	2.07	10.94	1.86	-5.45	3.82	0.33	2.75
4	43.0	16.81	4.20	5.86	1.31	-18.31	1.81	11.18	8.84
5	47.0	15.75	5.72	5.81	1.33	-18.22	2.04	9.69	8.95

	SCR RF joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau0 mean	tau 0 std	tau 1 mean	tau1 std
0	26.0	19.89	2.88	10.89	1.69	-5.19	4.77	17.40	5.58
1	26.0	21.46	3.74	10.35	1.61	-6.10	4.20	20.07	6.64
2	70.0	12.11	3.08	11.21	1.73	-4.67	4.67	2.13	3.41
3	37.0	20.46	2.76	6.03	1.21	-18.48	1.79	18.23	6.02
4	41.0	11.28	3.29	6.01	1.10	-18.32	2.08	1.25	3.57

	SCQR RF joint								
Subgroup	count	ALT mean	ALT std	time mean	time std	tau0 mean	tau 0 std	tau 1 mean	tau1 std
0	51.0	20.62	3.11	11.66	1.91	-4.22	4.50	17.45	5.71
1	40.0	13.21	2.25	13.01	1.32	-2.45	3.33	1.54	3.66
2	22.0	12.16	2.93	9.79	0.64	-6.40	3.12	1.70	3.14
3	41.0	20.44	2.93	6.37	1.43	-17.40	2.59	18.06	6.02
4	46.0	13.35	2.69	5.91	1.32	-18.08	2.63	2.97	4.09

Table S12: Subgroup characteristics from joint partitioning on heteroscedastic synthetic data (MOP-JCI).

	SCR CMGP separate on outcome 1						
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 1 mean	tau 1 std
1.0	18.0	0.65	0.26	1.25	0.65	16.92	6.15
2.0	61.0	0.81	0.50	-0.71	0.56	17.17	6.61
3.0	35.0	-0.06	0.17	0.05	0.87	12.43	4.05
4.0	11.0	-1.05	0.56	0.63	1.11	9.05	3.34
5.0	25.0	-1.20	0.70	-0.04	0.79	8.48	3.56

	SCR CMGP separate on outcome 2						
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 2 mean	tau 2 std
1.0	27.0	-0.21	1.03	1.36	0.52	2.80	0.67
2.0	32.0	-0.06	0.88	0.19	0.27	2.14	0.43
3.0	18.0	-0.02	0.90	-0.89	0.40	1.44	0.57
4.0	73.0	0.35	0.85	-0.55	0.77	0.74	0.30

	SCR RF separate on outcome 1						
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 1 mean	tau 1 std
1.0	40.0	1.12	0.42	0.20	0.95	19.19	7.18
2.0	87.0	-0.35	0.75	-0.12	0.91	11.83	4.40
3.0	23.0	-0.47	0.70	-0.10	0.82	8.56	2.56

	SCR RF separate on outcome 2						
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 2 mean	tau 2 std
1.0	150.0	0.02	0.94	-0.03	0.91	1.5	0.88

Table S13: Subgroup characteristics from separate partitioning on IHDP data (Baseline R2P).

	SCR CMGP joint								
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 1 mean	tau 1 std	tau 2 mean	tau 2 std
0	33.0	0.77	0.37	0.24	0.91	19.47	6.00	2.12	0.61
1	19.0	-0.23	0.23	0.21	0.86	13.13	3.67	2.14	0.69
2	24.0	-1.62	0.93	0.63	0.88	8.41	3.70	2.34	0.83
3	74.0	0.02	1.08	-0.38	0.91	12.00	5.16	0.82	0.38

	SCR RF joint								
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 1 mean	tau 1 std	tau 2 mean	tau 2 std
0	19.0	1.06	0.46	-0.10	0.80	19.49	7.74	1.25	0.75
1	38.0	0.99	0.42	0.14	1.25	19.24	5.40	1.80	1.12
2	39.0	-0.71	0.91	0.77	0.71	10.52	6.00	2.02	0.83
3	54.0	-0.26	0.56	-0.84	0.41	10.74	4.26	0.92	0.55

	SCQR RF joint								
Subgroup	count	nnhealth mean	nnhealth std	momage mean	momage std	tau 1 mean	tau 1 std	tau 2 mean	tau 2 std
0	54.0	0.98	0.36	-0.16	0.95	18.20	6.14	1.49	0.94
1	40.0	-0.27	0.38	0.40	1.04	13.25	4.43	2.25	0.89
2	32.0	-0.07	0.29	-0.60	0.82	11.17	3.98	0.78	0.35
3	24.0	-1.63	0.81	0.13	1.14	7.57	2.65	1.60	0.93

Table S14: Subgroup characteristics from joint partitioning on IHDP data (MOP-JCI).