S.C. Anoke, S.-L. Normand, C.M. Zigler

# Supplementary Materials
for
Approaches to Treatment Effect Heterogeneity in the Presence of Confounding

Contents

R code to reproduce the results from the simulation study in §4 of the manuscript ("Simulation Study 4") can be found at the following URL:

`https://github.com/sanoke/approachesTEH`

Appendix 1 Table 5 from Simulation Study 4

Table 5 Summary of denominators used to calculate the cell-specific averages visualized in Figure 1 as part of the simulation study in §4 of the manuscript; scenarios are defined in Table 2. "TE(1)" denotes the estimated subgroup with the smallest ATE. Note that the estimation procedure using GBM was the only procedure that yielded subgroups with an undefined treatment effect, thus it is data associated with this procedure that is presented below. The cell-specific averages presented for BART and FS were across all 1000 simulation iterations.

	Scenario A	**Scenario B**	**Scenario C**	**Scenario D**	**Scenario D***
TE(1)	1000	1000	1000	1000	1000
TE(2)	1000	1000	1000	1000	1000
TE(3)	1000	1000	1000	1000	1000
TE(4)	1000	1000	1000	1000	1000
TE(5)	1000	1000	1000	1000	1000
TE(6)	1000	1000	1000	1000	1000
TE(7)	1000	1000	1000	1000	1000
TE(8)	1000	1000	1000	977	985
TE(9)	1000	1000	1000	676	680
TE(10)	990	1000	992	108	105
undefined TE	10	0	8	892	895

In interpreting the columns of Table 5 above, recall the scenario definitions given in the manuscript:

Scenario A confounding and no effect modification

Scenario B effect modification and no confounding

Scenario C effect modification and confounding

Scenario D effect modification and confounding by effect modifiers

Scenario D* effect modification and nonlinear confounding by effect modifiers

Appendix 2 Full, Annotated Summary Image from Simulation Study 4 (Figure 5)

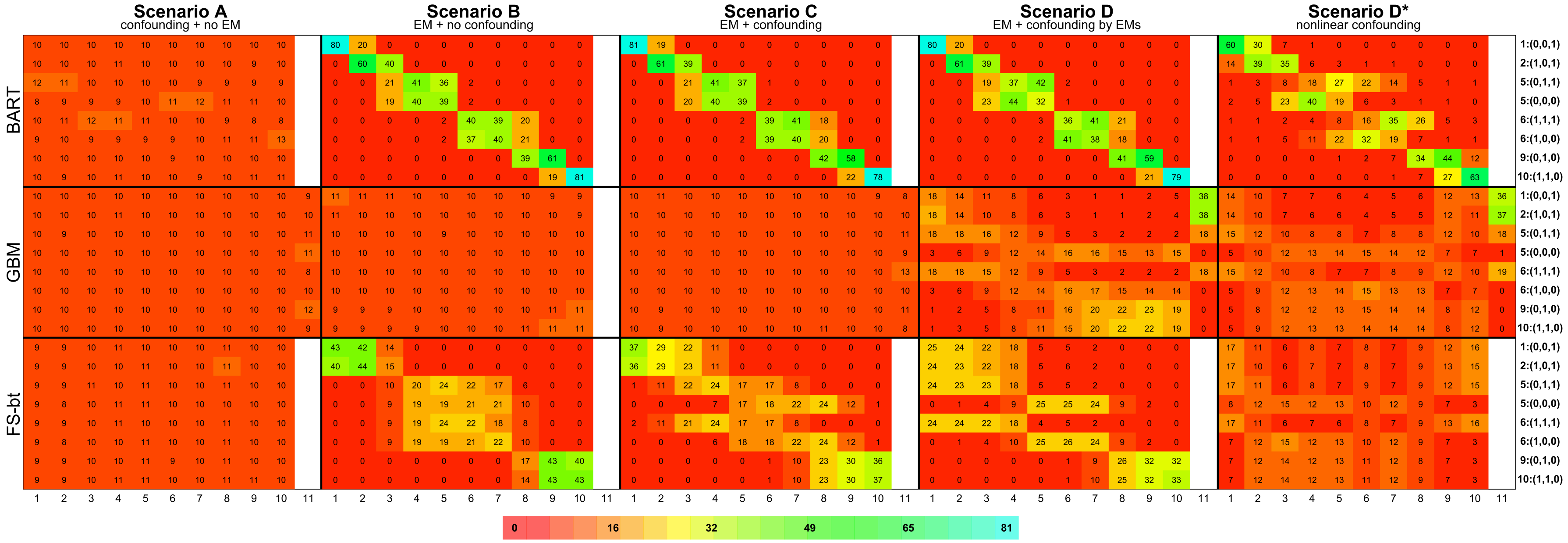

Figure 5 Annotated visualization of results from Simulation Study 4. This figure is an annotated version of Figure 1 in the main manuscript; each cell is annotated with the exact quantity that the underlying color represents, and each row is annotated with the corresponding subpopulation (in terms of the subpopulation average treatment effect and effect modifiers (E_1, E_2, E_3)). This figure also contains a visualization of the results from simulation scenario A, which was omitted from the manuscript for brevity. Details regarding how the figure was constructed, and how to interpret the figure, are given in §4.2 of the manuscript.

Appendix 3 Heuristic Biasedness Evaluation from Simulation Study 4 (Figure 6)

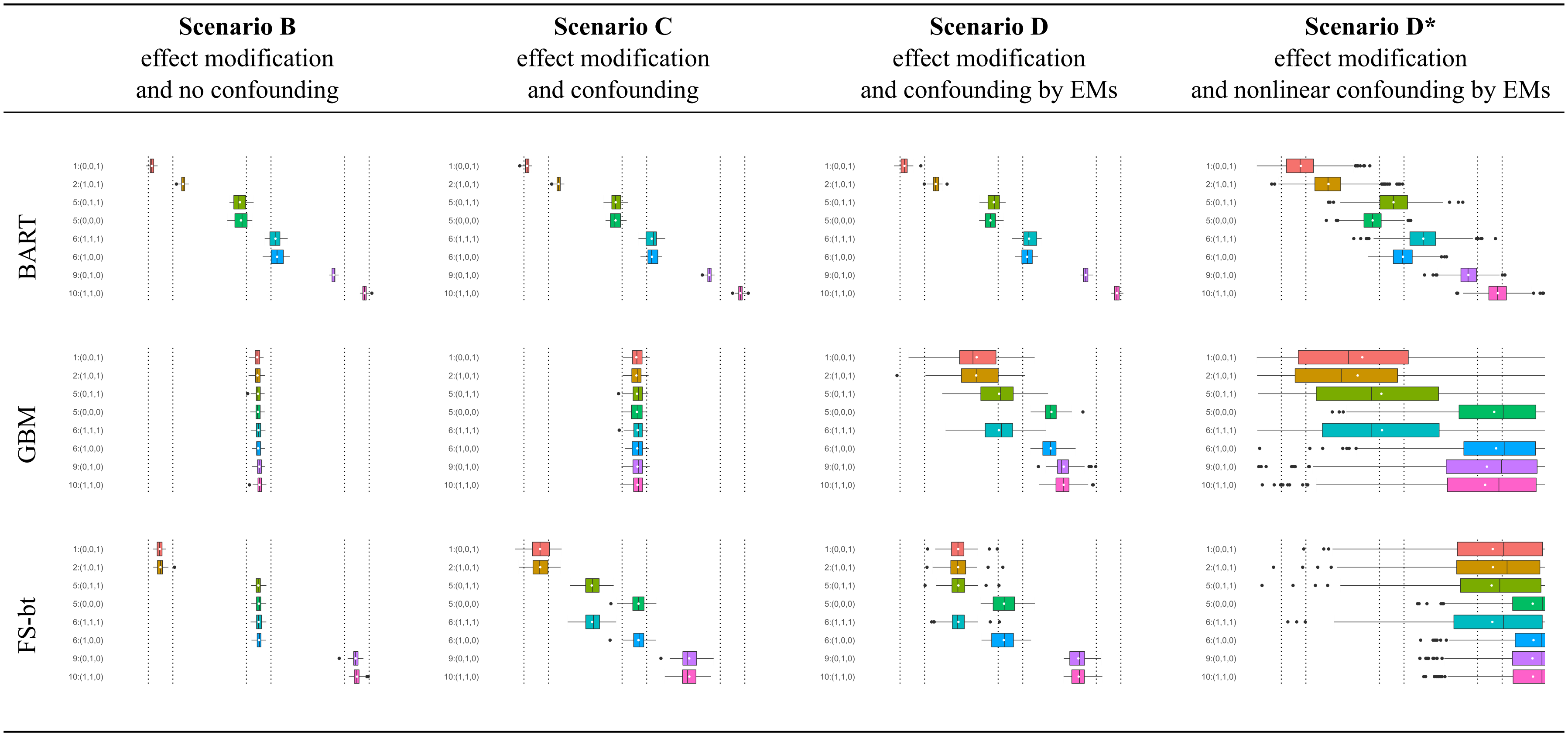

Figure 6 Forest plots by true treatment effect, from Simulation Study 4. The structure of this grid is pattered after Figure 1, where each row is an estimation method, and each column is a data generation scenario. To ease explanation, consider the forest plot associated with the BART analysis of data generated under Scenario A. Letting $j = 1, \ldots, 1000$ denote the simulation iteration, each observation during the jth iteration is assigned to a subgroup, within which an average treatment effect is estimated. Every observation in true Group 1 (say) has an associated ATE – the ATE estimated from the subgroup that the observation was assigned to. We can then take everyone in true Group 1, and take an average of these ATEs; in fact, we can do this for all eight true Groups, then plot the resulting eight averages. To generate the forest plots in this figure, these eight special averages were plotted, but for all 1000 simulation iterations. For each true subgroup, a boxplot is used to help visualize the distribution of estimates in that group. For clarity, the x-axes of the forest plots have been omitted, but there are vertical dashed lines denoting the true average treatment effects $(1, 2, 5, 6, 9, 10)$.

Appendix 3.1 Results for Simulation Scenario A (Figure 7)

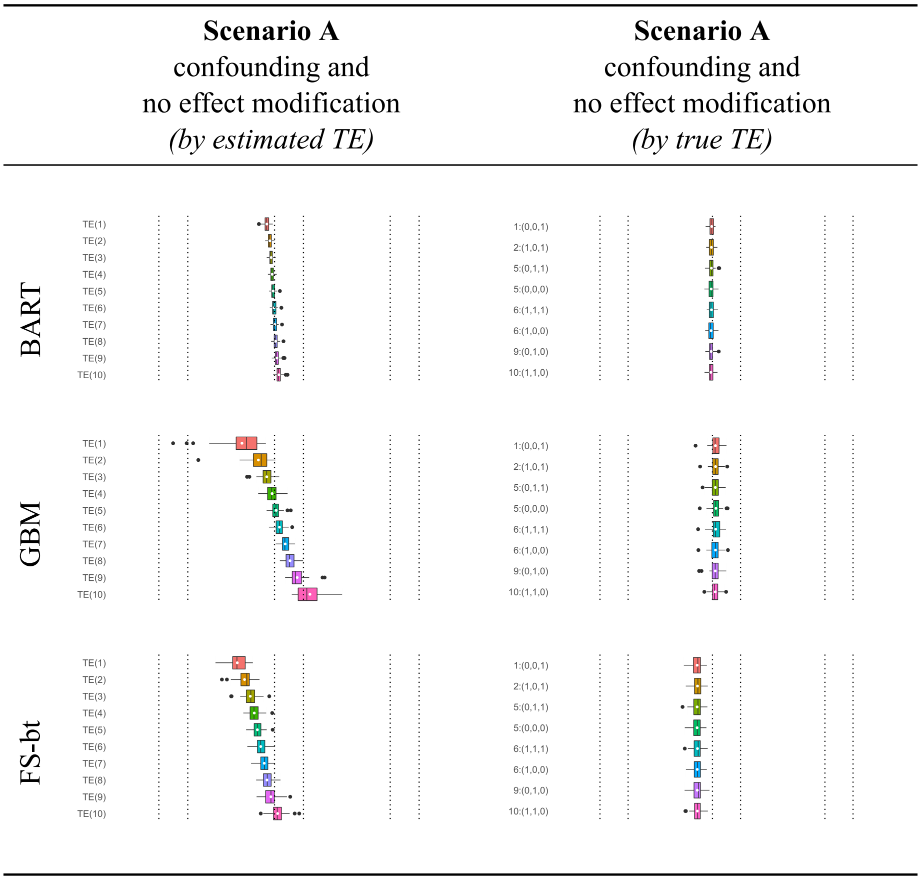

Figure 7 Forest plots by estimated and true treatment effect (respectively), under Simulation Scenario A from Simulation Study 4. For brevity, this scenario was not displayed within the main simulation results. For details on how to interpret the first or second column of this figure, see the caption associated with Figure 2 or Figure 6, respectively.

Appendix 4 Investigation of increased variance in Simulation Study 4 (Figure 8)

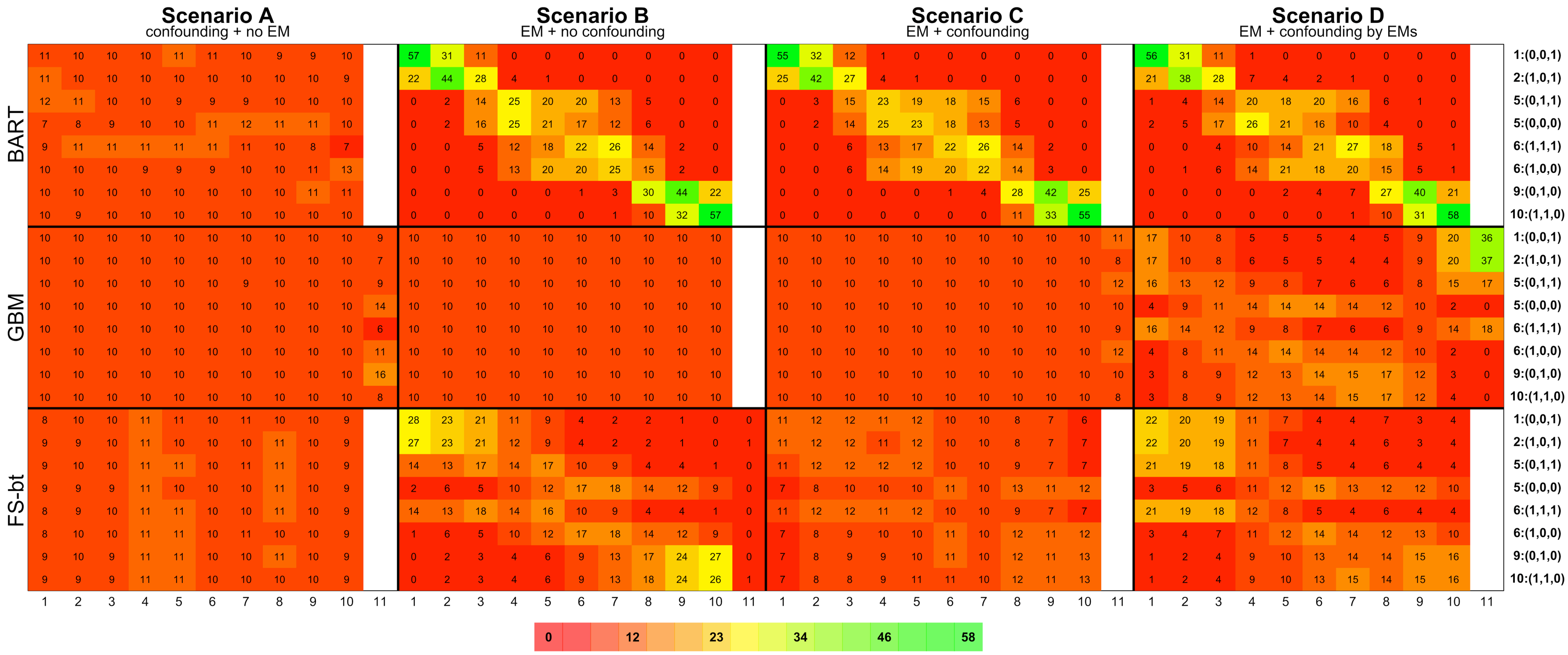

Figure 8 Annotated visualization of results from Simulation Study 4, where the variance is increased. This figure is a visual summary of a simulation study that is identical to the one discussed in §4 of the manuscript, with one major and one minor difference. The standard deviation has been increased from 1 unit to 10 units, to indicate less pronounced treatment effect heterogeneity (recall that the subpopulation-specific treatment effects for the eight subpopulations are $(1, 2, 5, 5, 6, 6, 9, 10)$). Also, this figure represents 100 rather than 1000 simulation iterations (the Facilitating Score simulation takes 12+ hours per iteration) although the results look the same under 1000. Comparing the resulting figure to the figure from the original simulation study (Figure 1 in the manuscript), we see that the same general patterns persist with BART and GBM, but the Facilitating Score shows some degradation in performance.

Appendix 5 Simulation Study: Simulated Treatment and Outcome (from real data parameters)

In this study, we consider the comparison of drug-eluting stents (DES) to bare-metal stents (BMS) as treatment of myocardial infarction (MI), by looking at the association of each with the two-year revascularization rate. We use real covariate data to simulate these two treatment options, as well as the two-year revascularization rate, to begin exploring the ability of these methods to identify TEH in real data.

Data Structure and Analysis

De-identified inpatient data on 38 covariates were generated by 169 539 Medicare beneficiaries hospitalized in the continental United States during 2009, 2010, or 2011 with their first MI. While the covariate summary given in Table 3 (in the original manuscript) is of hospitalizations in 2008, the 2009-2011 covariate distribution is very similar. As treatment, these patients underwent percutaneous coronary intervention (PCI) for the placement of exactly one type of stent, either a DES or BMS. Let $\boldsymbol{x}$ denote the covariate vector of a patient and $T \sim \text{Bern}(p_{\ell_2})$ their binary treatment indicator, where $p_{\ell_2} = \text{expit}(\boldsymbol{x}^\top \widehat{\boldsymbol{\alpha}})$ is the probability of receiving a DES. The value of coefficient $\widehat{\boldsymbol{\alpha}}$ was set as the maximum likelihood estimate from the regression of observed treatment on the 38 covariates. Binary outcome $Y \sim \text{Bern}(\mu_{\ell_2})$ indicates that the patient had been readmitted for revascularization (via CABG or a second PCI) within two years of discharge from their original MI hospitalization, or died before they could experience the revascularization event. It is modeled by setting $\mu_{\ell_2} = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} - 0.308T + 0.6T(elig) - T(diabetes)$, where $(\widehat{\boldsymbol{\beta}}, -0.308)$ is the maximum likelihood estimate from the regression of the observed outcome on $(\boldsymbol{x}, t)$, and $(0.6, -1)$ the fixed interaction coefficients for effect modifiers $elig$, an indicator of Medicaid eligibility, and $diabetes$, an indicator of prior diabetes diagnosis. The main effects of these covariates are contained in $\widehat{\boldsymbol{\beta}}$. These two covariates define four subgroups, with ATEs $(-0.22, -0.08, 0.02, 0.16)$ measured on the risk difference scale.

A dataset was simulated by first sampling the observed covariate data of 10 000 Medicare beneficiaries from the 169 539, then generating T and Y from the distributions described above. To preclude effect estimation issues caused by empirical positivity violations, any covariate with a prevalence of less than 5% in either treatment arm was dropped. To preclude obfuscation of our argument within this artificial simulation scenario, any dataset that did not include either effect modifier (i.e., after having been dropped for low prevalence) was discarded. This process was repeated to generate 100 simulated datasets, and each analyzed as described in §4.1 of the manuscript.

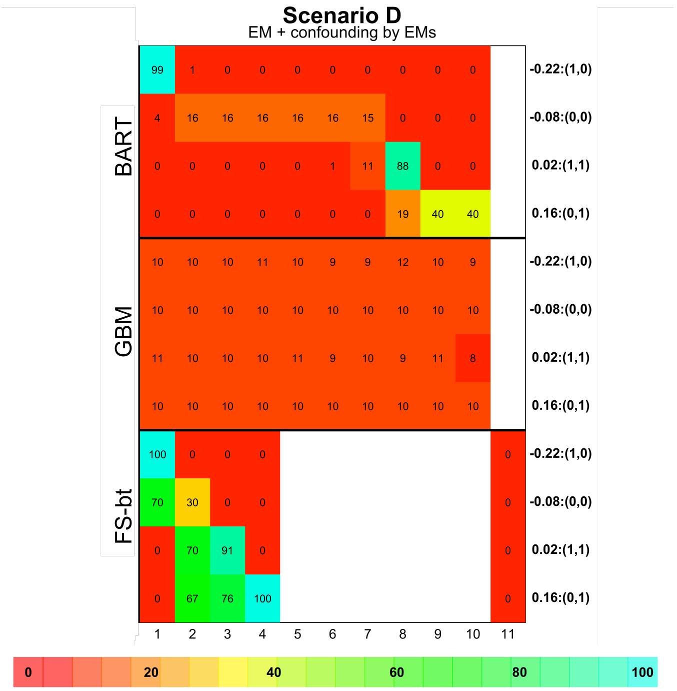

Figure 9 Visualization of results from this simulation study. Details regarding how the figure was constructed, and how to interpret the figure, are given in §4.1-4.2 of the manuscript (except there are 100 simulation iterations here).

Results and Discussion

Analysis results are summarized as described in §4.2 of the manuscript and visualized in Figure 9 of these Supplementary Materials (for one data generation scenario). Looking at the results generated by FS, we draw conclusions similar to those drawn in Simulation Study 4 in the main manuscript; namely, FS is able to detect effect modifiers that are strongly associated with either the outcome or treatment, where strength is relative to the associations of the other covariates to the treatment and outcome. In this example, prior diabetes diagnosis has a strong association with the outcome, and we see that FS is able to group observations by the value of this covariate: the top two rows have a prior diabetes diagnosis, and the bottom two rows do not.

An interesting point is the spread of color in the last two rows of the FS result block, which is caused by the particulars of PAM's estimation procedure. While we were able to prespecify that 10 subgroups be estimated, and the "center" of each of these subgroups is an observation from the dataset (by design), if no other observations are close to a selected center, then that center will remain in an estimated subgroup of size one. As applied to Figure 9 of these Supplementary Materials , PAM is typically able to detect 2-4 substantive subgroups. The remaining subgroups have just one observation, so have an undefined treatment effect; thus, empty columns 5 though 10 imply that PAM always had at least six estimated subgroups with one patient in each. The spread of color is actually an

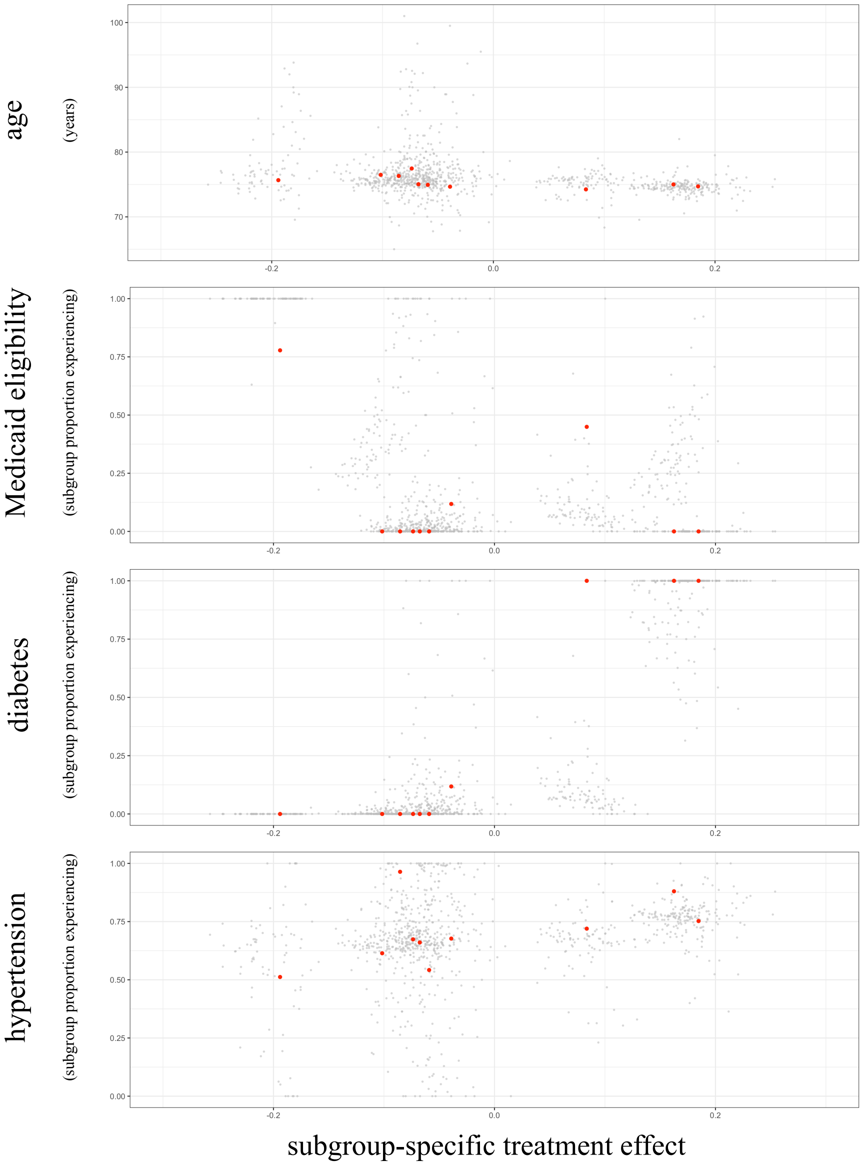

Figure 10 Visualization of results from a single dataset from this simulation study. Covariates displayed are *age*, *Medicaid eligibility*, prior *diabetes* diagnosis, and prior *hypertension* diagnosis. Details regarding how the figure was constructed, and how to interpret the figure, are described on page 10.

average across simulation iterations that detected differing numbers of subgroups.

Considering GBM, the results do not show evidence of being able to detect either EM, as demonstrated by the relative homogeneity of color and row percentages.

Finally considering the analysis using BART, this analysis procedure generated results that are as we would expect from correct identification of TEH. For example, the first row of this block represents observations in the subgroup with the smallest ATE, that are eligible for Medicaid and do not have a previous diabetes diagnosis, with a membership of $10\,000 \times 0.11 \times 0.73 = 803$. Because a decile is 1000 observations, we expect 100% of observations to be column 1, and this is what we see in Figure 9 (slightly different due to sampling variability). The remaining four rows also contain percentages as we expect.

A natural follow-up is to ask whether we can group individuals using their entire ITE posterior distribution, rather than a point summary, and Figure 10 of these Supplementary Materials attempts to answer this. Here we visualize one of the 100 datasets summarized in Figure 9 of these Supplementary Materials , to move our argument towards what we would expect in a real data analysis. To ease explanation, consider the top-most plot marked *age*. The y-axis represents the subgroup-specific average age in years, and the x-axis represents the subgroup-specific ATE (again measured on the risk-difference scale). A red dot represents a subgroup, generated as described earlier: the $10\,000$ posterior means are partitioned into deciles, and the average age and average TE (taken as the average of the posterior mean ITEs within that decile) are plotted. Thus, we expect ten red dots in this single plot. These red dots are clustered into four groups because, by design, there are four true subgroups. The number of red dots in these clusters is, as with the columns Figure 9, proportional to the size of the true subgroup (e.g., the largest true subgroup has an ATE of -0.08).

The posterior means of Figure 10 are somewhat redundant with Figure 9; it is the gray dots that provide the additional information on the full distribution of each observation's ITE. To generate the values that these gray dots represent, we applied the subgrouping process used on the posterior means (i.e., partitioned into ten subgroups and calculated subgroup-specific averages) to each of the 1000 posterior draws, and plotted a random subset of 100. We note that the same cutpoints were used to partition each posterior draw; namely, the deciles that were used to partition the posterior means. Thus from this figure we are able to visualize the posterior means in red, in addition to some measure of uncertainty in gray.

By design, *age* is not an EM, and does not present itself as such; its distribution remains approximately constant as the subgroup-specific treatment effect increases. However *diabetes* is an EM, where its presence is associated with a larger treatment effect, and we see this manifest as an upward trend within its plot. We see analogous patterning in the plot for *Medicaid eligibility*. Interestingly, *hypertension* was not *a priori* specified as an EM, but due to its positive correlation with diabetes (a known clinical phenomenon), displays itself as one: the subgroup-specific ATE increases as the prevalence of *hypertension* increases.