Abstract

We demonstrate how Hahn et al.’s Bayesian Causal Forests model (BCF) can be used to estimate conditional average treatment effects for the longitudinal dataset in the 2022 American Causal Inference Conference Data Challenge. Unfortunately, existing implementations of BCF do not scale to the size of the challenge data. Therefore, we developed \textit{flexBCF} — a more scalable and flexible implementation of BCF — and used it in our challenge submission. We investigate the sensitivity of our results to two \textit{ad hoc} modeling choices we made during our initial submission: (i) the choice of propensity score estimation method and (ii) the use of sparsity-inducing regression tree priors. While we found that our overall point predictions were not especially sensitive to these modeling choices, we did observe that running BCF with flexibly estimated propensity scores often yielded better-calibrated uncertainty intervals.
1 Introduction

Given pairs of observed predictor vectors $\mathbf{x} \in \mathbb{R}^p$ and scalar outputs $y \in \mathbb{R}$, Chipman et al. (2010)’s Bayesian additive regression trees (BART) model approximates the regression function $f(\mathbf{x}) := \mathbb{E}[y|\mathbf{x}]$ with a sum of binary regression trees. BART often delivers accurate point estimates and well-calibrated uncertainty intervals around evaluations $f(\mathbf{x})$ without requiring users to (a) pre-specify the functional form of $f$ or (b) tune any hyperparameters. BART’s ease-of-use and generally excellent tuning-free estimation, prediction, and uncertainty quantification have made BART a popular “off-the-shelf” modeling tool.

As Hill (2011) noted, these advantages make BART an attractive choice for estimating heterogeneous treatment effects. To wit, conditional average treatment effects may be non-linear and may depend on complicated interactions between confounders, covariates, and treatment. Correctly specifying such non-linearity and interactions is difficult — if not impossible — in most practical settings, making BART particularly compelling. Given observed triplets $(\mathbf{x}, y, z)$ of covariates $\mathbf{x}$, outcome $y$, and treatment indicator $z$, Hill (2011) used BART to obtain an estimate $\hat{f}(\mathbf{x}, z)$ of the response surface $\mathbb{E}[y|\mathbf{x}, z]$ before estimating treatment effects as $\hat{f}(\mathbf{x}, 1) - \hat{f}(\mathbf{x}, 0)$. In past iterations of the ACIC Data Challenge, BART-based models have consistently ranked among the top-performing methods. One method in particular — Hahn et al. (2020)’s Bayesian causal forests (BCF) model — stands out as particularly adept at estimating heterogeneous treatment effects. In this paper, we describe our experience deploying BCF to the 2022 ACIC Data Challenge.

At a high-level, BCF models the expected outcome as

$$\mathbb{E}[Y|\mathbf{x}, z] = \mu(\mathbf{x}) + z \times \tau(\mathbf{x})$$

and further approximates the prognostic function $\mu$ and the treatment effect function $\tau$ using separate regression tree ensembles. When fitting the tree ensembles, BCF includes an estimate of the propensity score as an additional covariate. Although these ensembles are \textit{a priori} independent, they are dependent \textit{a posteriori}, as BCF leverages all observations (both treated and control) to learn $\mu$.

At first glance, BCF appears ill-suited to the 2022 ACIC Data Challenge: BCF was initially proposed for cross-sectional data and the Data Challenge features longitudinal data with time-varying covariates and outcomes. However, as we show in Section 2.2, under a mild assumption about the data generating process, the Data Challenge admits an additive decomposition similar to the one in Equation (1). Also, while our analysis plan enables one to use BCF to estimate the treatment effect function — and subsequently the desired subgroup average treatment effects — we found that the implementation available in the \textsf{R} package \textsf{bcf} did not scale to the challenge data. We therefore re-implemented a version of BCF that can scale to the Data Challenge.
When writing our implementation and subsequently deploying it during the competition, we made a number of essentially *ad hoc* decisions. First, we included an estimate of Imai and Ratkovic (2014)’s covariate balancing propensity score (CBPS) as an additional covariate in the expected outcome model. Second, in our regression tree prior, we selected splitting variables using Linero (2018)’s sparsity-inducing Dirichlet prior instead of the uniform prior that is used in the *bcf* package. Briefly, Linero (2018)’s prior encourages regression trees to split on a small number of covariates, thereby facilitating a type of automatic variable selection. In the context of treatment effect estimation, such a sparsity-inducing prior can potentially identify the main drivers of effect heterogeneity. We assess the sensitivity of our contest submissions to these modeling choices. Specifically, we study the extent to which our results change when we use different estimates of the propensity score and when we use Chipman et al. (2010)’s uniform prior on the splitting variables in the trees.

Here is an outline for the rest of the paper. We briefly introduce the competition dataset in Section 2 before presenting our identification analysis and discussing our new implementation of BCF. Then, in Section 3, we compare the estimates for several configurations of the BCF model. Finally, we discuss the implications of our findings in Section 4.

## 2 Identification and estimation

In this section, we review the 2022 ACIC Data Challenge’s study design and identification strategy. We then outline our estimation strategy.

### 2.1 Review: Study design, notation, and definitions

The 2022 ACIC Data Challenge involved analyzing a total of 3,400 synthetic datasets consisting of 200 independent realizations of 17 different data generating processes. Each synthetic dataset simulated observations of beneficiary-level Medicare expenditure data over a period of four years. Each beneficiary received care at one of about 500 different medical practices. In addition to simulated time-invariant beneficiary- and practice-level covariates, each dataset also included average measurements of patient covariates within each practice. In each data generating process, treatment was assigned to some practices after two years. The main task of the Challenge was to assess how the treatment deployed at the practice level affected future medical expenditures of beneficiaries in the treated practices.

For each beneficiary *i* in practice *j*, let *Y*<sub>ijt</sub> denote their outcome (Medicare expenditure) in year *t*. Also, for each practice *j*, let *Z*<sub>j</sub> ∈ {0, 1} denote its treatment status where *Z*<sub>j</sub> = 1 indicates that practice *j* was assigned treatment between year 2 and year 3 and let *Z*<sub>j</sub> = 0 indicates that practice *j* was not treated during that period. As the notation suggests, the treatment is assigned only once, in between year 2 and year 3, and before assigning treatment, all beneficiaries are untreated.
Also, treatment is assigned at the practice level and all beneficiaries in a practice received the same treatment condition. Hence, we omit subscripts \( i \) and \( t \) in \( Z_j \) for notational simplicity\(^1\). Finally, for each beneficiary \( i \) in practice \( j \), let \( X_{ij} \in \mathbb{R}^p \) be all observed, “pre-treatment covariates,” i.e. variables measured at years 1 and 2 before treatment is assigned. These covariates can include time-invariant or time-varying variables in year 1 and 2, either at the practice level or at the beneficiary level. Again, for notational simplicity, we omit the subscript \( t \) in \( X_{ij} \)\(^2\).

We use potential outcomes to define causal effects (Splawa-Neyman et al., 1990; Rubin, 1974). For each beneficiary \( i \) in practice \( j \), let \( Y_{ijt}(1) \) denote their potential outcome at time \( t \) when their practice was treated between years 2 and 3. Similarly, let \( Y_{ijt}(0) \) denote their potential outcome at time \( t \) when their practice was untreated between years 2 and 3\(^3\).

We focus on the conditional average treatment effect on the treated (CATT), defined as

\[
\text{CATT}(x, t) = \mathbb{E}[Y_{ijt}(1) - Y_{ijt}(0)|Z_j = 1, X_{ij} = x], \quad t \in \{3, 4\}. \tag{2}
\]

Note that taking the average with respect to the distribution of the covariates \( X_{ij} \) gives us the overall average treatment effect on the treated (ATT), i.e. \( \text{ATT}(t) = \mathbb{E}[\text{CATT}(X_{ij}, t)] \). For both estimands (and contrary to the Data Challenge), the expectation is taken over a hypothetical super-population of practices and beneficiaries at each time \( t \) in order to better align with the Bayesian paradigm we pursue in the paper\(^4\).

\(^1\)We can also define treatment status to be \( Z_{jt} \in \{0, 1\} \) where \( Z_{jt} = 1 \) if practice \( j \) was treated in year \( t \) and \( Z_{jt} = 0 \) if practice \( j \) was not treated in year \( t \). We can then encode the aforementioned constraints on the treatment status by assuming: (i) for every practice \( j \) and year \( t = 1, 2, \mathbb{P}(Z_{jt} = 0) = 1 \) and (ii) for every practice \( j \), \( \mathbb{P}(Z_{jt} = Z_{j(t-1)}) = 1 \).

\(^2\)Again, we can also define pre-treatment covariates by partitioning them into time-varying variables and time-invariant variables, similar to the notation used to define causal effects in longitudinal settings (Hernan and Robins, 2020, Chapter 19). Specifically, for each beneficiary \( i \) in practice \( j \), let \( X_{ijt} \) denote their vector of time-varying variables in year \( t = 1, 2 \). We also let \( X_{ij0} \) denote their vector of time-invariant variables. Then, we can let \( X_{ij2} = (X_{ij0}, X_{ij1}, X_{ij2}) \) be the collection of all pre-treatment covariates up to (and including) year 2.

\(^3\)The potential outcomes notation makes some implicit assumptions, notably (i) treatment status of practice \( j \) cannot affect the potential outcome of beneficiaries in practice \( j' \neq j \), (ii) there are no different versions of treatment (or non-treatment), (iii) treatment status does not vary amongst beneficiaries in the same practice, and (iv) treatment status does not change once assigned in between years 2 and 3. Conditions (i) and (ii) are often known as the Stable Unit Treatment Value Assumption (Rubin, 1980, 1986).

\(^4\)The super-population we have in mind is based on “model-based” sampling (Särndal, 1978) where for each time \( t \), the entire collection of random variables \( (Y_{11t}(1), Y_{11t}(0), Z_1, X_{11}, \ldots, Y_{n_j,t}(1), Y_{n_j,t}(0), Z_j, X_{n_j,t}) \) follows some distribution \( F \) and the expectation is over the distribution \( F \). For example, suppose at each time \( t \), we have \( (Y_{ijt}(1), Y_{ijt}(0), Z_j, X_{ij}) \overset{i.i.d.}{\sim} G \) where \( G \) can be a parametric distribution or a mixture of parametric distributions, say a mixture of \( J \) parametric densities where each density represents one of the \( J \) practices. Then \( F = \prod_{ij} G \) and the expectation \( \mathbb{E}[Y_{ijt}(1) - Y_{ijt}(0)|Z_j = 1, X_{ij} = x] \), which is over \( F \), is equivalent to the expectation over \( G \). Historically, model-based sampling has been associated with Bayesian paradigms as it allows investigators to encode subject-matter expertise by imposing constraints on \( F \); see, e.g., Ericson (1969), Särndal (1978, §5), Gelman and Hill (2006, Chapter 21.2), and and Schochet (2013) for further discussions. Under appropriate conditions, an estimator (or a test) of the super-population estimand is also reasonable for estimating the finite-sample analog of the estimand; see Lehmann and Romano (2005, pg. 188) for one example and Ding et al. (2017).
2.2 Review: Causal Identification

To identify the CATT from the observed data, the Data Challenge made the following three assumptions for all beneficiaries and practices:

(A1) For all $t$, $Y_{ijt} = Y_{ijt}(0) \times 1 \ (t \leq 2) + Y_{ijt}(Z_j) \times 1 \ (t > 2)$.

(A2) For each $s \in \{1, 2\}$ and $t \in \{3, 4\}$, $(Y_{ijt}(0) - Y_{ijt}(s)) \perp \perp Z_j \mid X_{ij}$.

(A3) There is a $\delta > 0$ such that $P(Z_j = 1) > \delta$ and $P(Z_j = 1 \mid X_{ij} = x) < 1 - \delta$ for all $x$.

Under assumptions (A1)–(A3), for all $s = \{1, 2\}$ and $t \in \{3, 4\}$, we can write $\text{CATT}(x, t)$ as a function of the observed data using standard arguments from difference-in-differences (e.g. Heckman et al., 1997; Abadie, 2005):

$$
\text{CATT}(x, t) = \mathbb{E}[Y_{ijt} \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijt} \mid Z_j = 0, X_{ij} = x] - \left\{ \mathbb{E}[Y_{ij1} \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ij1} \mid Z_j = 0, X_{ij} = x] \right\}.
$$

(3)

We provide a full proof in Appendix A. Additionally, using Equation (3) “twice” under different values of $s = 1, 2$, we observe

$$
\mathbb{E}[Y_{ij2} - Y_{ij1} \mid Z_j = 1, X_{ij} = x] = \mathbb{E}[Y_{ij2} - Y_{ij1} \mid Z_j = 0, X_{ij} = x]
$$

(4)

In words, Equation (4) roughly states that conditional on covariates, the outcome trend before year 3 is identical between the treated and the untreated groups. This additional structure implied by the assumptions from the Data Challenge is the basis for our modeling assumption in Section 2.3 that eventually allows us to utilize a faster version of BCF.

2.3 Estimation & Implementation

We will additionally assume that for every $(t, x, z)$, we can write

$$
\mathbb{E}[Y_{ijt} \mid X_{ij} = x, Z_j = z] = \mu(x, t) + z \times 1 \ (t > 2) \times \tau(x, t).
$$

(5)

Under Equation (5) and the assumptions introduced in Section 2.2, we have $\text{CATT}(x, t) = \tau(x, t)$. Also under Equation (5), we have $\mathbb{E}[Y_{ij2} - Y_{ij1} \mid Z_j = z, X_{ij} = x] = \mu(x, 2) - \mu(x, 1)$ for $z \in \{0, 1\}$, satisfying Equation (4). In contrast to Assumption (A2), which was about trends in outcomes, Equation (5) makes an assumption directly about the outcome at each time $t$. A benefit of our additional assumption is that the functional forms of Equations (1) and (5) are now similar, making it possible to use BCF to estimate the CATT. We note that BCF allows $\mu$ and $\tau$ to depend on different covariates. We included the average of the patient-level covariates at both pre-treatment
time points as covariates for \( \mu \) and similar averages at all four time points as covariates for \( \tau \). Although patient-level covariates are time-invariant, practice composition generally changed over time as patients attrited from the study. Consequently, our conditioning on these post-treatment covariates can lead to biased estimation of the CATT unless patient attrition is not impacted by treatment (Rosenbaum, 1984). For our initial submission and in our subsequent re-analysis, we assumed that patient attrition — and therefore practice composition — was not impacted by treatment.

Unfortunately, we encountered several problems when using the \texttt{bcf} package to fit the model in Equation (5) to the beneficiary-level data. First, the implementation in that package begins by internally copying all of the covariate data into a new matrix and allocating a matrix to hold posterior samples of evaluations of \( \tau \) for all observations. When modeling the beneficiary-level data, we included the practice label as an additional categorical covariate. \texttt{bcf} converted the practice label into several hundred binary indicators (one per practice). Consequently, \texttt{bcf} allocated around 5GB of memory to store a copy of each beneficiary-level dataset. Further, \texttt{bcf} attempted to allocate around 20GB of memory for a matrix storing 2000 posterior samples of the effect of treatment on each beneficiary. These memory requirements were well beyond the scope of our personal machines. When we attempted to run BCF on the full data on a high-memory computer, we found that the sampler was extremely slow. On further inspection, we discovered that the posterior updates of each regression tree involved several redundant computations. Basically, while updating a single tree, \texttt{bcf} loops over the entire dataset several times to identify the leaf to which every observation is assigned. Because the underlying Gibbs sampler only changes at most two leaves at a time, the map from observations to tree leaves does not change much iteration to iteration, rendering many of these loops redundant.

We developed a more efficient implementation of BCF that (i) avoids making unnecessary copies of the data; (ii) does not make redundant calculations when updating the trees; (iii) does not instantiate a matrix to hold all posterior draws of \( \tau(x, t) \); and (iv) provides users the option to use Linero (2018)'s sparsity-inducing prior on the splitting variables used by each tree. Our sampler returns a list of character vectors whose elements contain string representations of each regression tree in the ensemble used to approximate \( \tau \). We wrote a new predict routine that takes these character vectors to output posterior samples of evaluations of \( \tau \). Our implementation is available in an \texttt{R} package that we call \texttt{flexBCF}. Compared to \texttt{bcf}, \texttt{flexBCF} is much faster and much less memory-intensive: our analyses ran in a few hours on shared computers in a high-throughput computing cluster and we needed only about 4GB of memory to analyze each beneficiary-level dataset with \texttt{flexBCF}. 

6
3 Re-analysis of Challenge Data

In our submission to the ACIC Data Challenge, we ran flexBCF using a sparsity-inducing prior on splitting variables and with an estimated covariate balancing propensity score. These choices were, admittedly, rather *ad hoc*. To see whether our submitted results change substantially under different choices, we re-analyzed the competition data using different propensity score estimation methods and using (or not) the sparsity-inducing prior on splitting variables. In addition to the covariate-balancing propensity score estimate (hereafter CBPS), we considered propensity scores estimated with BART (hereafter BART), $L_1$-regularized logistic regression (hereafter LASSO), and gradient boosting (hereafter GBM). We computed the four propensity score estimates using the R packages CBPS (Fong et al., 2022), BART (Sparapani et al., 2021), glmnet (Simon et al., 2011), and GBM (Greenwell et al., 2022). Note that CBPS and LASSO assume that the log-odds of treatment are linear in the covariates $x$ while BART and GBM do not make any *a priori* assumptions about the functional form of the log-odds of treatment. Accordingly, we refer to CBPS and LASSO as parametric and BART and GBM as non-parametric propensity score estimators. We will refer to the model used in our initial Challenge submission as CBPS(S) where the suffix (S) refers to the sparsity-inducing prior. We will use the suffix (D) to refer to the uniform prior on splitting variables.

We re-analyzed the practice-level data using all eight configurations of propensity score and sparsity-inducing prior and we re-analyzed the beneficiary-level data using only CBPS(S), CPBS(D), GBM(S), and GBM(D). We compared the root mean square error (RMSE) and two uncertainty interval diagnostics (coverage and relative interval length to CBPS(S)) for (i) the overall ATT (i.e. the estimated CATT’s averaged over all beneficiaries and practices) and (ii) twelve subgroups effects defined as the estimated CATT averaged over different pre-specified subgroups defined by pre-treatment practice covariates. Throughout, we report these performance measures within several categories of DGPs defined by the amount of confounding and impact heterogeneity. Specifically, we consider five categories of DGPs: (i) no confounding but large impact heterogeneity; (ii) weak confounding with small impact heterogeneity; (iii) weak confounding with large impact heterogeneity; (iv) strong confounding with small impact heterogeneity; and (v) strong confounding with large impact heterogeneity.

3.1 Practice level results

**Overall ATT.** Figure 1 compares the RMSEs of all methods for the overall ATT for each DGP broken down by category. In general, as the amount of confounding and treatment effect heterogeneity increased, the RMSE of all methods increased. Further, we did not observe substantial differences between the methods within each category of DGP. Interestingly, in the absence of confounding, using a nonparametric propensity score estimate (i.e. BART or GBM) yielded slightly
larger RMSE than using parametric estimates (CBPS or L1-regularized logistic regression). However, in the presence of even weak confounding, the nonparametric propensity score estimates fared slightly better. That said, the difference in performance is quite small and is almost certainly within the range of Monte Carlo variability. Additionally, using the sparsity-inducing prior did not seem to improve the overall RMSE.

For each of the 17 DGPs, we computed the coverage of the 90% posterior credible interval of the ATT. Figure 2 reveals that all methods tended to have lower than nominal coverage, with coverage decreasing as the amount of confounding and effect heterogeneity increased. In the presence of confounding, the use of parametric propensity score estimates appeared to yield diminished coverage.

Figure 3 compares the lengths of the 90% posterior credible intervals produced by each method relative to CBPS(S). Values larger than one indicate that the method yielded longer uncertainty intervals. Figures 1–3 suggest that in the presence of confounding, fitting BCF with flexible propensity score estimate (i.e. BART or GBM) might not improve estimation but could yield wider uncertainty intervals with much better frequentist coverage than parametric propensity score estimates (i.e. CBPS or LASSO).

**Subgroup effects.** We similarly compared the estimation performance and uncertainty quantification on the eight models for each subgroup effect. For brevity, we present performance results averaged over all twelve subgroups in the main text and defer subgroup-by-subgroup comparisons to
Figure 2: Coverage of the 90% posterior credible interval for the overall ATT for each DGP category using practice-level data. Note the unconfounded category contains only one DGP while the other four categories contain 4 DGPs.

the Appendix B. Figure 4 compares the RMSE averaged across all twelve subgroup effects. Similar to the overall ATT, we did not observe substantial differences in the quality of estimated subgroup effects across the different methods.

Further, we found that nonparametric propensity score estimators yielded slightly longer posterior credible intervals that displayed higher frequentist coverage than parametric propensity score estimators. We observed similar results for each subgroup; see Figures 9–14 in Appendix B.

To summarize, the practice level analyses do not indicate a clear preference for any of the eight models we compared in terms of point estimation. However, we did observe that more flexible propensity score estimates (i.e. BART and GBM) had better-calibrated uncertainty intervals than parametric propensity score estimates (i.e. CBPS and LASSO). Finally, the use of sparsity-inducing priors did not seem to affect performance much at all.

3.2 Beneficiary-level results

In this section we compare the performance of four of methods fit to the full beneficiary-level data.

Overall ATT. Figure 5 shows the RMSE for the overall ATT across all beneficiaries in each category of DGP for the dense and sparse models based on a CBPS- and GBM-based propensity score estimate.

In the absence of confounding, there is very little discernible difference between the RMSE for
Figure 3: Lengths of 90% posterior credible interval for the overall ATT relative to CBPS(S) for each DGP category using practice-level data.

Figure 4: RMSE averaged across the twelve subgroup effects for each DGP category using practice-level data.

the four models. Similar to findings in the previous subsection, in the presence of confounding, the GBM-based models have somewhat lower RMSE as compared to the CBPS based models. The RMSE for the beneficiary level models are slightly higher when compared to the practice level counterparts in the presence of confounding, but are essentially the same when there is no
Figure 5: RMSE for overall ATT for each DGP category using beneficiary-level data

confounding.

Figures 6 and 7 compare the coverage and relative lengths of the 90% posterior credible intervals for the ATT. As was the case in the practice level models, using a CBPS-based propensity score estimate yielded shorter intervals with worse coverage than using a GBM-based propensity score estimate.

Figure 6: Coverage for overall ATT for each DGP category using beneficiary-level data.
Interestingly, the uncertainty intervals in the beneficiary level analysis have much lesser coverage as compared to practice level versions of the same models. The difference in coverage is quite large in the presence of confounding.

Subgroup effects. Figure 8 shows boxplots of the RMSE averaged across the 12 subgroup effects for each category of DGP. We found that the RMSE for the subgroup effects are relatively insensitive to the choice of propensity score estimate; see also Figures 15 and 16 in Appendix B.2 for comparisons of performance for each subgroup.

Interestingly, we found that when we averaged the RMSEs over the 12 different subgroup, using the sparsity-inducing prior yielded slightly smaller average RMSE. Although using the sparsity-inducing prior also tended to produce somewhat shorter uncertainty intervals, the difference in relative interval lengths was quite small; see Figures 17–20 in Appendix B.

4 Discussion

In this work, we showed that this year’s ACIC Data Challenge admitted a decomposition similar to the one used by Hahn et al. (2020)’s Bayesian casual forest model. Because the bcf package could not scale to the size of the data in the competition, we implemented a faster and less resource-intensive version, which is available in the flexBCF package. We investigated the sensitivity of our competition submission to two modeling choices - (i) the choice of propensity score method, and (ii) enforcing (or not) sparsity in the BART priors, for both practice-level and beneficiary-level
modeling. We found that the beneficiary-level modeling did not give any more accurate results than the practice level modeling. Although we did not observe much sensitivity to our modeling choices on the Challenge datasets, our re-analysis does suggest that fitting BCF-like models with flexible propensity score estimates may yield slightly longer uncertainty intervals with better frequentist coverage.

Acknowledgements

We are grateful to Mariel Finucane, Jennifer Starling, and Dan Thal from Mathematica Policy Research for several helpful discussions following the Data Challenge.

This research was performed using the computing resources and assistance of the UW–Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW–Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the OSG Consortium, which is supported by the National Science Foundation and the U.S. Department of Energy’s Office of Science.

Support for A.H.K. and S.K.D. was provided by the University of Wisconsin–Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.
References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Ding, P., Li, X., and Miratrix, L. W. (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2).

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society: Series B*, 31(2):195–233.

Fong, C., Ratkovic, M., and Imai, K. (2022). *CBPS: Covariate Balancing Propensity Score*. R package version 0.23.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2022). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.1.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If?* CRC Press.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A*, 147(5):656–666.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):575–582.

Rubin, D. B. (1986). Comment. *Journal of the American Statistical Association*, 81(396):961–962.

Särndal, C.-E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5(1):27–52.

Schochet, P. Z. (2013). Estimators for clustered education rcts using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics*, 38(3):219–238.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1):1–66.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.
Appendix

A Identification

By assumption (A3), we have $\delta < \mathbb{P}(Z_j = 1) < 1 - \delta$ so that the population consists of a non-trivial mixture of treated and untreated practices. Also, there are some values of $x$ where $\delta < \mathbb{P}(Z_j = 1 \mid X_{ij} = x) < 1 - \delta$.

Let $t$ be a post-intervention time-point (i.e. $t \in \{3, 4\}$) and let $s$ be a pre-intervention time-point (i.e. $s \in \{1, 2\}$). Then, for any $s$ and $t$ where $\delta < \mathbb{P}(Z_j = 1 \mid X_{ij} = x) < 1 - \delta$, we have the following:

$$
\begin{align*}
\mathbb{E}[Y_{ijt} \mid Z_j = 1, X_{ij} = x] &- \mathbb{E}[Y_{ijt} \mid Z_j = 0, X_{ij} = x] \\
= \mathbb{E}[Y_{ijt}(1) \mid Z_j = 1, X_{ij} = x] &- \mathbb{E}[Y_{ijt}(0) \mid Z_j = 0, X_{ij} = x] \\
= \mathbb{E}[Y_{ijt}(1) - Y_{ijt}(0) \mid Z_j = 1, X_{ij} = x] \\
&+ \mathbb{E}[Y_{ijt}(0) \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijt}(0) \mid Z_j = 0, X_{ij} = x] \\
= \text{CATT}(x, t) \\
&+ \mathbb{E}[Y_{ijt}(0) \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijt}(0) \mid Z_j = 0, X_{ij} = x] \\
= \text{CATT}(x, t) \\
&+ \mathbb{E}[Y_{ijt}(0) - Y_{ijs}(0) \mid X_{ij} = x] - \mathbb{E}[Y_{ijt}(0) - Y_{ijs}(0) \mid X_{ij} = x] \\
= \text{CATT}(x, t) \\
&+ \mathbb{E}[Y_{ijs}(0) \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijs}(0) \mid Z_j = 0, X_{ij} = x] \\
= \text{CATT}(x, t) + \mathbb{E}[Y_{ijs}(0) \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijs}(0) \mid Z_j = 0, X_{ij} = x] \\
= \text{CATT}(x, t) + \mathbb{E}[Y_{ijs} \mid Z_j = 1, X_{ij} = x] - \mathbb{E}[Y_{ijs} \mid Z_j = 0, X_{ij} = x]
\end{align*}
$$

Rearranging we can express the estimand of interest in terms of expectations of observables as
follows:

\[
\text{CATT}(\mathbf{x}, t) = \mathbb{E}[Y_{ijt} \mid Z_j = 1, \mathbf{X}_{ij} = \mathbf{x}] - \mathbb{E}[Y_{ijt} \mid Z_j = 0, \mathbf{X}_{ij} = \mathbf{x}]
\]

\[
- \left\{ \mathbb{E}[Y_{ij} \mid Z_j = 1, \mathbf{X}_{ij} = \mathbf{x}] - \mathbb{E}[Y_{ij} \mid Z_j = 0, \mathbf{X}_{ij} = \mathbf{x}] \right\},
\]

which is the same as Equation (3).

B Additional figures

In Section B.1, we present the subgroup effects i.e. CATT’s for both the practice-level and beneficiary-level models. Since covariates \( X_1, X_3, \) and \( X_5 \) are binary, we present their results together, and do the same for the categorical covariates \( X_2 \) and \( X_4 \). Figures 9 and 10 present the RMSE for the CATTs defined by binary and categorical covariates respectively; Figures 11 and 12 show the frequentist coverage of the uncertainty intervals for the CATTs defined by binary and categorical covariates respectively, while Figures 13 and 14 show interval lengths relative to the \( \text{CBPS(S)} \) model.

We follow the same template for the beneficiary-level modeling results in Section B.2: Figures 15 and 16 show the RMSE for the beneficiary-level model CATT estimates, Figures 17 and 18 show the frequentist coverage and Figures 19 and 20 present the relative interval lengths of the uncertainty intervals (on a logarithmic scale) obtained from the beneficiary-level model.
B.1 Practice level modeling

Figure 9: RMSE for binary covariates’ CATT for each DGP category using practice-level data
Figure 10: RMSE for categorical covariates’ CATT for each DGP category using practice-level data
Figure 11: Interval coverage for binary covariates’ CATT for each DGP category using practice-level data
Figure 12: Interval coverage for categorical covariates’ CATT for each DGP category using practice-level data
Figure 13: Relative interval length for binary covariates’ CATT for each DGP category using practice-level data
Figure 14: Relative interval length categorical covariates’ CATT for each DGP category using practice-level data
B.2 Beneficiary-level modeling

Figure 15: RMSE for binary covariates’ CATT for each DGP category using beneficiary-level data
Figure 16: RMSE for categorical covariates’ CATT for each DGP category using beneficiary-level data
Figure 17: Interval coverage for binary covariates’ CATT for each DGP category using beneficiary-level data
Figure 18: Interval coverage for categorical covariates' CATT for each DGP category using beneficiary-level data.
Figure 19: Relative interval length for binary covariates’ CATT for each DGP category using beneficiary-level data. Note: The lengths are displayed on a logarithmic scale.
Figure 20: Relative interval length categorical covariates’ CATT for each DGP category using beneficiary-level data. Note: The lengths are displayed on a logarithmic scale.