APPENDIX F. The BART Prior

The priors on the functions $m(\cdot)$ and $\tau(\cdot)$ can be indicated as (Chipman et al., 2010):

$$\sum_{r=1}^{R} g(\boldsymbol{x}_i; T_r, M_r) \tag{F1}$$

where T is a Bayesian regression/classification tree (B-CART) as described in Chipman et al. (1998) with b terminal nodes, each with an associated parameter $\mu_b \in M = \{\mu_1, \dots, \mu_B\}$. These trees partition the covariate space $\boldsymbol{x}_i$ via binary splits of the single components $x_{ip} \in \{x_{i1}, \dots, x_{iP}\}$. As with 'traditional' CART algorithms (Breaiman et al., 1984), $g(\boldsymbol{x}_i; T, M)$ defines a function that assigns $\mu_b \in M$ to $\boldsymbol{x}_i$. In other words, for a hypothetical model $Y_i = g(\boldsymbol{x}_i; T, M) + \varepsilon$, $E[Y_i|\boldsymbol{x}_i]$ corresponds to the node parameter $\mu_b \in M$ assigned through $g(\boldsymbol{x}_i; T, M)$. Similarly, for an ensemble of R trees as the one indicated in equation (F1), the function $g(\boldsymbol{x}_i; T_r, M_r)$ assigns the leaf parameter $\mu_{br} \in M_r$ to the corresponding observation vector $\boldsymbol{x}_i$, while the resulting conditional expectation is given by the sum $\sum_{r=1}^{R} \mu_{br}$.

Equation (F1) has three essential parameters (or parameter sets), R, T_r and M_r, to which we can add the standard deviation of ε_i in equation (5), σ. Keeping R fixed, the resulting sum-of-trees model is fully determined by the set $\{(T_1, M_1), \dots, (T_R, M_R), \sigma\}$ which controls all the terminal node parameters, the structure of the trees and the splitting decision rules. To each component of this parameter set, Chipman et al. (2010) impose a regularizing prior, which keeps the contribution of each tree relatively low within the ensemble. This is given by the joint distribution:

$$\begin{aligned}\phi((T_1, M_1), \dots, (T_R, M_R), \sigma) &= \left[\prod_{r=1}^{R} \phi(T_r, M_r|\sigma)\right] \phi(\sigma) \\ &= \left[\prod_{r=1}^{R} \phi(T_r, M_r)\right] \phi(\sigma)\end{aligned}$$

$$= [\prod_{r=1}^{R} \phi(M_r|T_r)p(T_r)]\,\phi(\sigma)$$

The above formulation postulates independence across trees, as well as independence between each tree and σ. Finally, conditional independence across the terminal node is also imposed, leading to:

$$\phi((T_1, M_1), \ldots, (T_R, M_R), \sigma) = \{\prod_{r=1}^{R} [\prod_{b=1}^{B} \phi(\mu_{br}|T_r)]\,\phi(T_r)\}\,\phi(\sigma)$$

The prior on T_r

Following Chipman et al. (1998), the prior on the tree structure of each single B-CART in the ensemble has three hyperparameter sets: (i) the distribution governing the splitting variable assignment at each interior node; (ii) the distribution governing the splitting rule at each interior node, conditional on the variable selected in (ii); (iii) the probability that some node with depth $d \in [0, \infty)$ will be nonterminal. Chipman et al. (1998) suggests using uniform distributions for both (i) and (ii), while (iii) can be indicated as:

$$\phi(d) = \alpha(1 + d)^{-\beta}$$

where $\alpha \in (0,1)$ and $\beta \in [0, \infty)$. The authors suggest a range of sensible values for α and β, each resulting in priors generating very shallow or moderately shallow trees. In our work, we calibrate α and β via a training-validation procedure, where the grid of candidates for the two hyperparameters is defined as in Chipman et al. (1998-2010): (1) $(\alpha = 0.5, \beta = 0.5)$ with mean $d \approx 2.1$ (very simple tree), (2) $(\alpha = 0.95, \beta = 0.5)$ with mean $d \approx 7$ ('bushy' tree) and $(\alpha = 0.95, \beta = 2)$ with mean $d \approx 3$ (simple tree).

The prior on $\mu_{br}|T_r$

Chipman et al. (2010) propose conjugate normal prior $\mathcal{N}(0, s_\mu^2)$ on $\mu_{br}|T_r$. Upon rescaling of the dependent variable such that $\bar{y} = 0$, $\min(y) = -0.5$ and $\max(y) = 0.5$ (this is performed

automatically by the `bcf` and `bartCause` packages in `R`), this prior is specified so that it assigns high probability to the interval $(-0.5,0.5)$. This is achieved by choosing $m_\mu$and s_μ such that $Rm_\mu - k\sqrt{R}s_\mu = -0.5$ and $Rm_\mu + k\sqrt{R}s_\mu = 0.5$. However, because the above transformation involves centering y, it suffices to define:

$$\phi(\mu_{br}|T_r) = \mathcal{N}(0, 0.5k^{-1}R^{-1/2})$$

where k can be calibrated via train-validation or cross-validation. Notice that as R or k grow, the prior on μ_{br} becomes tighter, thereby imposing heavier shrinkage on the terminal node parameters. This is a desirable property, as it prevents individual trees from playing an overwhelming role as the ensemble is formed. Our grid of candidates for k includes: (1) $k = 1$ (low shrinkage) and (2) $k = 3$ (high shrinkage).

The prior on σ

The prior Chipman et al. (2010) suggest for σ^2 is a conjugate inverse-chi square distribution:

$$\phi(\sigma^2) = \text{Inv}\chi^2(\lambda, \nu)$$

where λ indicates the scale of the distribution, while ν represent the degrees of freedom of the corresponding χ^2 distribution. Our grid of candidates for λ and ν is based on the data-driven approach proposed by the authors: given a sample estimate of $\sigma, \hat{\sigma}$, λ is set so that $P(\sigma < \hat{\sigma}) = \lambda$, i.e.: the λ^{th} quantile of σ corresponds to $\hat{\sigma}$. Therefore, sensible choices for λ include 0.75, 0.90 and 0.99. On the other hand, ν is free to vary between 3 and 10 to guarantee that the resulting prior has an appropriate shape.

Choosing R

When BART are used to perform estimation of prediction, R represents yet another hyperparameter in the model. Chipman et al. (2010) and Hahn et al. (2020) suggest a default value of $R = 200$, which

is described as striking a good balance between computation time and prediction accuracy. However, our testing shows that for the dataset used in this paper, 200 trees tend to provide unstable predictions over repeated estimations. On possible explanation for this behaviour might be related to the large number of predictors which, conditional on the optimized hyperparameters discussed so far, do not allow for stable representation of the unknown conditional expectation. Since the above-referenced papers do not provide clear guidelines for setting R, we refer to the standard literature on random forests (Breiman 2001; Friedman et al., 2009; Wager et al., 2014) and basic understanding of the modelling approach to set the size of the ensemble. As Chipman et al. (2010) state "[…] as R is increased, starting with $R = 1$, the predictive performance of BART improved dramatically, until at some point it levels off and very slowly degrades […] Thus, for prediction, it seems only important to avoid choosing R too small.". The authors also explain that decreasing R could diminish the redundancy of the predictors in the sum-of-trees model, which would help to explain the instability or our predictions at $R = 200$. Turning to the frequentist literature, we notice that Stetter et al. (2022) may have a similar issue when using generalized random forests (Athey et al., 2019), which may explain their choice of setting $R = 5000$. Increasing the number of trees to obtain stable estimates is also discussed in Wager et al. (2014). Therefore, we make several attempts to look for what value of R guarantees a satisfying stability of our estimates. Ultimately, we find that $R = 2000$ provides for good results in that predictions remain stable across multiple estimations and the algorithm does not become prohibitively slow.

Calibration results

To sum up, our grid of candidates for the prior hyperparameters discussed so far is shown in Table F1 with $R = 2000$. Our calibration proceeds as follows: (i) we split the dataset in two parts, obtaining two subsamples: training set (70% of the observations) and validation set; (ii) we fit a standard BART algorithm under configuration j, where $j \in \{1, \dots, 18\}$, to the training set (this is done using the `R`

package `bartCause`); (iii) we evaluate the predictive accuracy of the resulting ensemble using the validation set and select the parameter set that guarantees the best predictive performance. We next use this configuration to estimate model (7), except we adjust the prior for $\tau(\boldsymbol{x}_i)$ to impose a stronger regularization, as suggest by Hahn et al. (2020). Specifically, given predictively optimal hyperparameters $\alpha_{opt}, \beta_{opt}, \nu_{opt}, \lambda_{opt}$ and k_{opt} we set $\alpha^{\tau}_{opt} = \frac{\alpha_{opt}}{3}$, $\beta^{\tau}_{opt} = 2\beta_{opt}$, $\nu^{\tau}_{opt} = \nu_{opt}$, $\lambda^{\tau}_{opt} = \lambda_{opt}$ and $k^{\tau}_{opt} = \frac{k_{opt}}{2}$, which roughly reflects the recommended defaults used in the `bcf` package. Therefore, the final configuration we use in the present paper is summarized in Table F2.

Table F1. *Grid of hyperparameter candidates for the BART algorithm*

Configuration	α	β	ν	λ	k
1	0.5	0.5	3	0.9	1
2	0.5	0.5	3	0.9	3
3	0.5	0.5	3	0.99	1
4	0.5	0.5	3	0.99	3
5	0.5	0.5	10	0.75	1
6	0.5	0.5	10	0.75	3
7	0.95	0.5	3	0.9	1
8	0.95	0.5	3	0.9	3
9	0.95	0.5	3	0.99	1
10	0.95	0.5	3	0.99	3
11	0.95	0.5	10	0.75	1
12	0.95	0.5	10	0.75	3
13	0.95	2	3	0.9	1
14	0.95	2	3	0.9	3
15	0.95	2	3	0.99	1
16	0.95	2	3	0.99	3
17	0.95	2	10	0.75	1
18	0.95	2	10	0.75	3

Table F2. *Hyperparameters for the BCF in Equation (7)*

Treatments	α_{opt}	β_{opt}	ν_{opt}	λ_{opt}	k_{opt}
T2	0.95	0.5	3	0.99	3
T1	0.95	0.5	10	0.75	3

References

Athey, S., J. Tibshirani, and S. Wager. 2019. "Generalised Random Forests." *Annals of Statistics* 47 (2): 1148–78.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Chipman, H. A., E. I. George, and R. E. McCulloch. 1998. "Bayesian CART Model Search." *Journal of the American Statistical Association* 93 (443): 935–48.

———. 2010. "Bart: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4 (1): 266–98.

Hahn, P. R., J. S. Murray, and C. M. Carvalho. 2020. "Bayesian Regression Tree Models for Causal Inference: Regularisation, Confounding, and Heterogeneous Effects (With Discussion)." *Bayesian Analysis* 15 (3): 965–1056.

Wager, S., T. Hastie, and B. Efron. 2014. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife." *Journal of Machine Learning Research* 15 (1): 1625–51.