

A Appendix

A.1 Causal inference

Confounder A confounder is a variable that is associated with both the treatment and the outcome, causing a spurious correlation. For instance, summer is associated with eating ice cream and getting sunburns, but there is no causal relationship between the two.

Propensity score model A propensity score model is a function that predicts treatment from the observed covariates i.e. $P(T = 1|C = c)$ for a binary treatment T and a covariate vector C .

Potential outcome As defined by the Rubin causal model (39), a potential outcome $Y(t)$ is the value that Y would take if T were set by (hypothetical) intervention to the value t .

Identification assumptions Inference is possible under three identification assumptions.

- **No interference** For a given individual i , this assumption implies that $Y_i(t)$ represents the value that Y would have taken for individual i if T had been set to t for individual i , i.e the potential value of Y_i if T_i had been set to t .
- **Consistency** For a given individual i , $T_i = t \Rightarrow Y_i = Y_i(t)$. This means that for individuals who actually received treatment level t , their observed outcome is the same as what it would have been had they received treatment level t via an hypothetical intervention.
- **Conditional exchangeability** For a given individual i , we assume that conditional on C , the actual treatment level T is independent of each of the potential outcomes:
 $Y(t) \perp T \mid C, \forall t$

A.2 Evaluation of the tree consistency

We evaluated the consistency of our clustering across subsamples using an Adjusted Rand Index (33). Given a set of n elements $S = \{o_1, \dots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \dots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \dots, Y_s\}$, a partition of S into s subsets, define the following:

- a , the number of pairs of elements in S that are in the same subset in X and in the same subset in Y .
- b , the number of pairs of elements in S that are in different subsets in X and in different subsets in Y .
- c , the number of pairs of elements in S that are in the same subset in X and in different subsets in Y .
- d , the number of pairs of elements in S that are in different subsets in X and in the same subset in Y .

The Rand index, R , is:

$$R = \frac{a + b}{a + b + c + d} = \frac{a + b}{\binom{n}{2}}$$

Intuitively, $a + b$ can be considered as the number of agreements between X and Y and $c + d$ as the number of disagreements between X and Y . The adjusted Rand index is the corrected-for-chance version of the Rand index (40).

A.3 Further related work

This section was mistakenly referred to as section 3.1 in the main text, 218.

Causal inference provides a wide range of methods for estimating causal effect from data with unbalanced treatment allocation. In balancing methods such as matching or weighting methods, the data is pre-processed to create subgroups with lower treatment imbalance or “natural experiments”.

Matching Matching methods consist of clustering similar units from the treatment and control groups to reduce imbalance. In general, a matching procedure generates weights w_{ij} denoting the assignment of one or many control units j to a treated unit i ((41), Chapter 5). Exact matching only assigns control units to treatment units with the exact same set of covariate values. But typically, matched control units j are chosen based on a nearest neighbours search according to some distance metric. However, matching procedures induce a bias-variance trade-off as discarding unmatched

samples reduces estimation error at the cost of increased variance. However, all matching methods suffer from the curse of dimensionality, making them impractical in high-dimensional datasets. Exact matching and coarsened exact matching (42) find exponentially fewer matches as the input dimension grows (43). Alternative methods include Propensity score matching (44), where distance is computed from an estimate of the propensity score $P(T = 1 | X = x)$, and Mahalanobis distance matching (28) (see more details below in 4.1). However, compression into a single dimension can lead to highly unrelated matches with very different characteristics in the original covariate space, and can ultimately increase estimation bias (45). *Clivio et. al* overcome this issue by developing a multivariate balancing scores to perform matching for high-dimensional causal inference. Nevertheless, this approach is not interpretable.

Weighting methods An alternative to matching are *weighting* methods, where sample weights are estimated, generalising the problem formulation of matching. In Inverse Probability Weighting (IPW) (4) which is the most popular alternative in that category, the samples are weighted according to their *propensity* score, i.e. the estimated probability of treatment conditional on their covariates.

Adjustment methods *Adjustment* methods estimate the causal effect from regression outcome models where both treatment and covariates act as predictors of the outcome. These regressions can be fitted through various methods like linear regression (5), neural networks (6; 7), or tree-based models (8). Common alternatives include Doubly Robust estimators (46), Double Debiased Machine Learning (47) and metalearners such as the T-learner and X-learner (9). These methods have the advantage of being very data-adaptive. More particularly, the Causal Tree approach (8) builds on regression tree methods, and splits the data to optimize for goodness of fit in treatment effects. Causal Tree separates the training dataset into two subsamples: a splitting subsample and an estimating subsample. The splitting subsample is used to build a causal tree while the estimating subsample is used to generate unbiased conditional treatment effect estimates. This procedure is called “honest estimation” and is anticipated to avoid overfitting.

A.4 Experimental details: synthetic datasets

A.4.1 Natural experiment dataset

For the natural experiment dataset, we consider a Death outcome D , a binary treatment of interest T and two covariates such that $X = (S, A)$ with S the sex and A the continuous age such that:

$$S \sim \text{Bernoulli}(0.5) \\ A \sim \text{Normal}(50, 20^2)$$

The sample size was chosen to be $N = 20,000$.

We defined four sub-populations, each constituting a natural experiment, with a different propensity distribution $P(T = 1 | X = x)$:

$$\begin{aligned} \Pr(T = 1 | S = 1, A \geq 50) &\sim \text{TruncatedNormal}_{[0,1]}(0.5, 0.1^2) \\ \Pr(T = 1 | S = 1, A < 50) &\sim \text{TruncatedNormal}_{[0,1]}(0.3, 0.1^2) \\ \Pr(T = 1 | S = 0, A \geq 50) &\sim \text{TruncatedNormal}_{[0,1]}(0.1, 0.1^2) \\ \Pr(T = 1 | S = 0, A < 50) &\sim \text{TruncatedNormal}_{[0,1]}(0.4, 0.1^2) \end{aligned}$$

Individual treatment propensities were sampled from the corresponding distributions above and observed treatment values were sampled from a Bernoulli distribution parameterized with the individual propensities. The outcome probabilities were not modeled as a distribution. Instead, observed outcome values were sampled directly from a Bernoulli distribution parameterized with a constant value that depended on both X and T .

- $\Pr(Y | T = 1, S = 1, A \geq 50) \sim \mathcal{B}(0.1)$
- $\Pr(Y | T = 1, S = 1, A < 50) \sim \mathcal{B}(0.2)$
- $\Pr(Y | T = 1, S = 0, A \geq 50) \sim \mathcal{B}(0.4)$
- $\Pr(Y | T = 1, S = 0, A < 50) \sim \mathcal{B}(0.15)$
- $\Pr(Y | T = 0, S = 1, A \geq 50) \sim \mathcal{B}(0.2)$
- $\Pr(Y | T = 0, S = 1, A < 50) \sim \mathcal{B}(0.4)$
- $\Pr(Y | T = 0, S = 0, A \geq 50) \sim \mathcal{B}(0.8)$
- $\Pr(Y | T = 0, S = 0, A < 50) \sim \mathcal{B}(0.3)$

No positivity violation was modeled in this experiment.

635 A.4.2 Positivity violations dataset

For this second synthetic experiment, we build a dataset with two positivity violating subgroups. Let us consider a synthetic example of a dataset with a Death outcome D , a binary treatment of interest T and three binary covariates –sex, cancer and arrhythmia– such that $X = (S, C, A)$. For the marginal distributions, we set:

$$S \sim \text{Ber}(0.5)$$

$$C \sim \text{Ber}(0.3)$$

$$A \sim \text{Ber}(0.1)$$

636 The sample size was chosen to be $N = 20,000$.

- 637 • $\Pr(T = 1|S = 1, C = 1, A = 1) \sim \text{TruncatedNormal}_{[0,1]}(1.00, 0.02^2)$
- 638 • $\Pr(T = 1|S = 1, C = 0, A = 1) \sim \text{TruncatedNormal}_{[0,1]}(0.32, 0.10^2)$
- 639 • $\Pr(T = 1|S = 1, C = 1, A = 0) \sim \text{TruncatedNormal}_{[0,1]}(0.12, 0.10^2)$
- 640 • $\Pr(T = 1|S = 1, C = 0, A = 0) \sim \text{TruncatedNormal}_{[0,1]}(0.42, 0.10^2)$
- 641 • $\Pr(T = 1|S = 0, C = 1, A = 0) \sim \text{TruncatedNormal}_{[0,1]}(0.17, 0.10^2)$
- 642 • $\Pr(T = 1|S = 0, C = 1, A = 1) \sim \text{TruncatedNormal}_{[0,1]}(0.30, 0.10^2)$
- 643 • $\Pr(T = 1|S = 0, C = 0, A = 1) \sim \text{TruncatedNormal}_{[0,1]}(0.24, 0.10^2)$
- 644 • $\Pr(T = 1|S = 0, C = 0, A = 0) \sim \text{TruncatedNormal}_{[0,1]}(0.00, 0.02)$

645 The observed outcome values were sampled from a Bernoulli distribution parameterized with a
646 constant value that depended on both X and T . For the treated:

- 647 • $\Pr(Y|T = 1, S = 1, C = 1, A = 1) \sim \mathcal{B}(0.13)$
- 648 • $\Pr(Y|T = 1, S = 1, C = 1, A = 0) \sim \mathcal{B}(0.08)$
- 649 • $\Pr(Y|T = 1, S = 1, C = 0, A = 1) \sim \mathcal{B}(0.21)$
- 650 • $\Pr(Y|T = 1, S = 1, C = 0, A = 0) \sim \mathcal{B}(0.1)$
- 651 • $\Pr(Y|T = 1, S = 0, C = 1, A = 1) \sim \mathcal{B}(0.36)$
- 652 • $\Pr(Y|T = 1, S = 0, C = 1, A = 0) \sim \mathcal{B}(0.29)$
- 653 • $\Pr(Y|T = 1, S = 0, C = 0, A = 1) \sim \mathcal{B}(0.24)$
- 654 • $\Pr(Y|T = 1, S = 0, C = 0, A = 0) \sim \mathcal{B}(0.09)$

655 For the untreated:

- 656 • $\Pr(Y|T = 0, S = 1, C = 1, A = 1) \sim \mathcal{B}(0.31)$
- 657 • $\Pr(Y|T = 0, S = 1, C = 1, A = 0) \sim \mathcal{B}(0.4)$
- 658 • $\Pr(Y|T = 0, S = 1, C = 0, A = 1) \sim \mathcal{B}(0.29)$
- 659 • $\Pr(Y|T = 0, S = 1, C = 0, A = 0) \sim \mathcal{B}(0.45)$
- 660 • $\Pr(Y|T = 0, S = 0, C = 1, A = 1) \sim \mathcal{B}(0.4)$
- 661 • $\Pr(Y|T = 0, S = 0, C = 1, A = 0) \sim \mathcal{B}(0.51)$
- 662 • $\Pr(Y|T = 0, S = 0, C = 0, A = 1) \sim \mathcal{B}(0.43)$
- 663 • $\Pr(Y|T = 0, S = 0, C = 0, A = 0) \sim \mathcal{B}(0.73)$

664 A.5 Further experiment results: synthetic experiments

665 A.5.1 Natural experiment dataset

666 **Partitioning** The tree partitions so to recreate the four intended sub populations where a natural
667 experiment was simulated. The average Rand index was equal to 0.901 across the 50 subsamples
668 ($\sigma = 0.263$). Violating leaf nodes are marked in red.

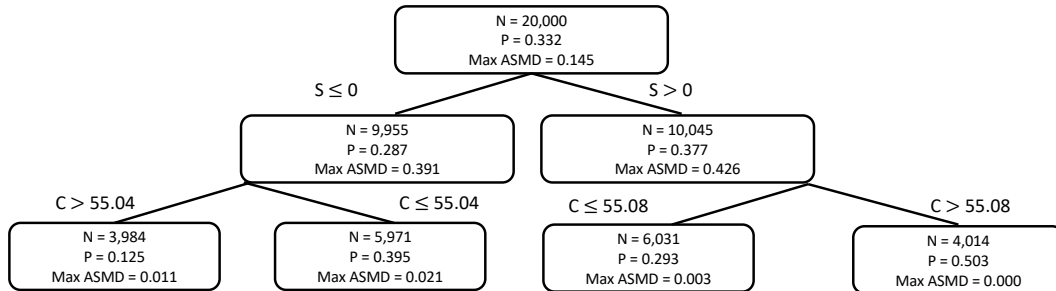


Figure A1: Tree structure after training on the entire natural experiment dataset ($N = 20,000$). Violating leaf nodes are marked in red.

669 **Causal effect estimation** Here, Causal Tree has both higher estimation bias and higher estimation
670 variance compared to other methods. The tree-like nature of our data structure may be incompatible
671 with the optimization function of Causal Tree, which maximizes on treatment effect heterogeneity.
672 Causal Forest, which has multiple Causal Tree models, however overcomes this difficulty and
673 performs well.

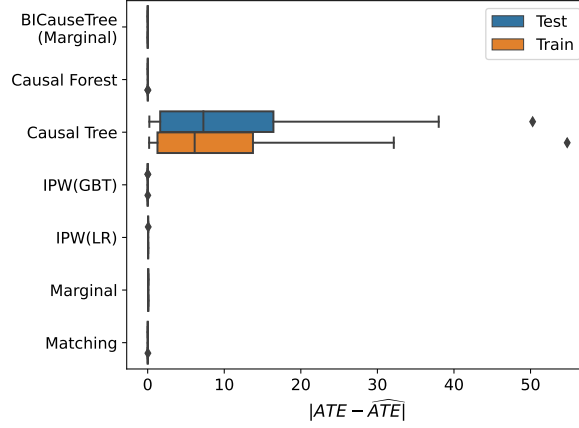


Figure A2: Estimation bias for the natural experiment dataset across 50 subsamples with $N = 20,000$.

674 **Propensity score estimation** BICauseTree(Marginal)’s propensity estimation is well calibrated.
675 It is closer to the identity line than logistic regression- (IPW(LR)) and the gradient boosting trees-
676 (IPW(GBT)) based models. The calibration however remains satisfactory across all models. This
677 further shows our ability to identify natural experiments in a simple data setting.

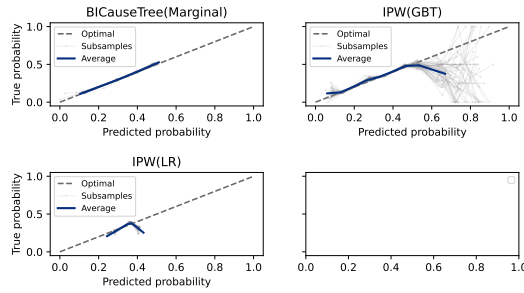


Figure A3: Calibration of the propensity score estimation for the natural experiment dataset across 50 subsamples, on the natural experiment testing set ($N = 10,000$).

678 **Outcome estimation** The performance of BICauseTree w.r.t outcome estimation is only satisfactory
679 in this experiment. This shows our ability to estimate the ATE despite a somehow lower calibration
680 of our outcome estimation. Ultimately, the performance would likely have been improved had we
681 used a more complex outcome model. The calibration of the predicted outcomes by BICauseTree
682 however remains better than the one by Matching.

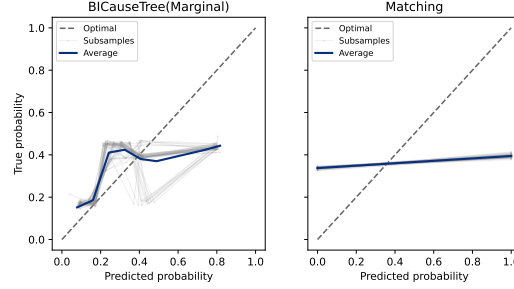


Figure A4: Calibration of the outcome estimation for the natural experiment dataset across 50 subsamples on the natural experiment testing set ($N = 10,000$).

683 **Effect estimation bias reduction with tree depth** Figure A5 shows how estimation bias decreases
684 as we increase the maximum depth hyperparameter of our *BICauseTree(Marginal)* on the natural
685 experiment training dataset ($N = 10,000$). Here, each circle in the plot represents the effect
686 estimated for each subsample. **Note that the circles were mistakenly described as corresponding to**
687 **nodes instead of subsamples in the main text, on line 301.** The dotted line shows the average bias
688 with an IPW estimator. The shaded area represents the 95% confidence interval (CI) for IPW. This
689 result shows that our estimation is robust to the choice of a maximum tree depth hyperparameter.
690 Above a certain threshold, the effect estimate from *BICauseTree* remains unbiased.

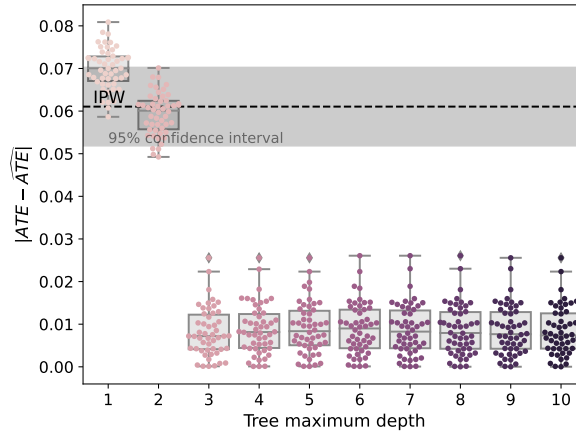


Figure A5: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the natural experiment training set ($N = 10,000$) across 50 subsamples.

691 **Treatment allocation bias reduction with tree depth** Figure A6 below shows the weighted
692 ASMD of both covariates S and A in *BICauseTree* models with varying maximum tree depth
693 hyperparameters. It illustrates the reduction of treatment allocation bias as the tree grows, but also
694 shows that this reduction is a heuristic as the ASMD might not decrease monotonically.

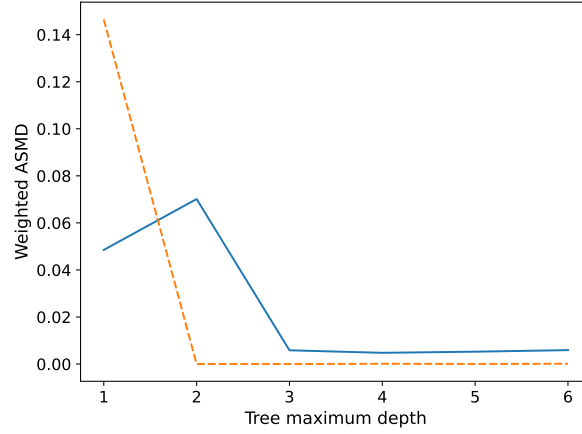


Figure A6: Weighted ASMD for all covariates applying *BICauseTree(Marginal)* models with varying maximum tree depths on the natural experiment training set ($N = 10,000$) across 50 subsamples.

695 A.5.2 Positivity violations dataset

696 **Partitioning** BICauseTree coherently flags the two subpopulations where a positivity violations
 697 were simulated, as shown in the example partition on the entire dataset below in Figure A7 (violating
 698 leaves are marked in red). The average Rand index was equal to 0.986 across the 50 subsamples
 699 ($\sigma = 0.026$).

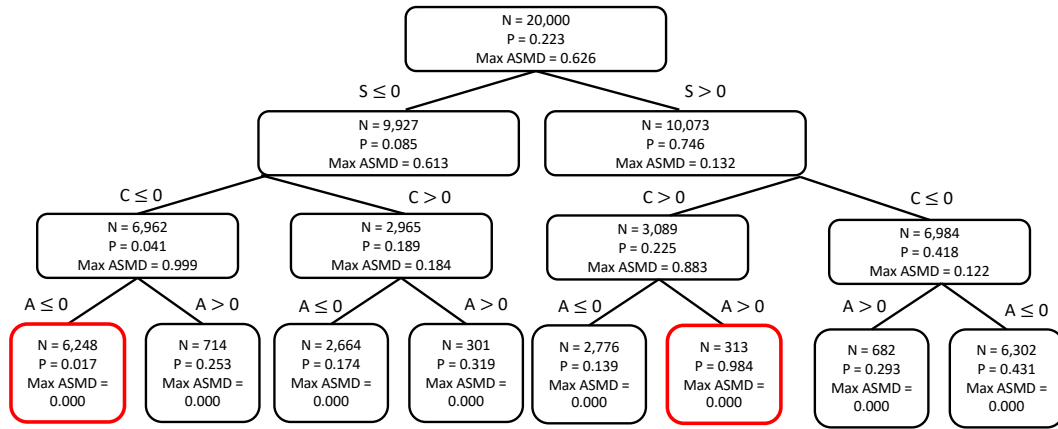


Figure A7: Tree structure after training on the entire positivity violations dataset ($N = 20,000$). Violating leaf nodes are marked in red.

700 **Causal effect estimation** Our tree compares with all existing alternatives. Matching has both
 701 higher estimation bias and higher estimation variance compared to other methods.

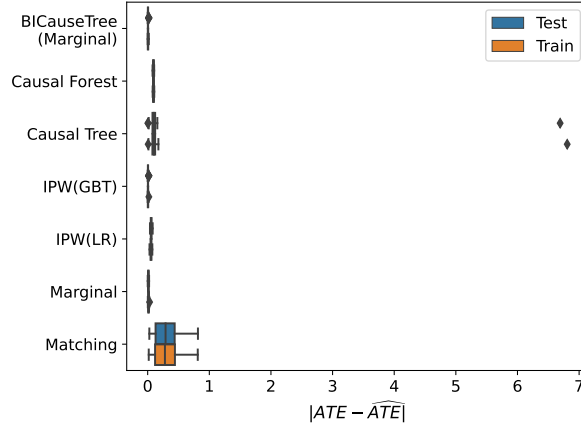


Figure A8: Estimation bias on the positivity violations dataset ($N = 20,000$) across 50 subsamples

Propensity score estimation BICauseTree(Marginal) shows good calibration, since it indeed clearly identified the correct subpopulations. The gradient-boosting tree (IPW(GBT)) also shows good calibration, as the tree-based modeling captures the underlying structure of the data. On the other hand, the logistic regression-based model (IPW(LR)) is ill-specified for the data and therefore presents relatively poorer calibration.

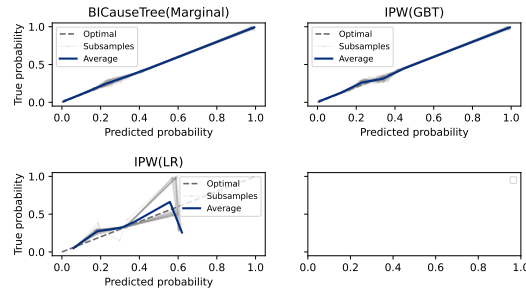


Figure A9: Propensity score calibration comparing our approach to existing alternatives on the testing set of the positivity violation dataset ($N = 10,000$) across 50 subsamples

Outcome estimation BICauseTree shows descent calibration on outcome prediction in this experiment. Matching shows poor calibration.

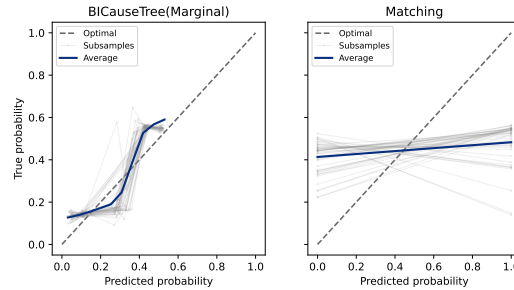


Figure A10: Calibration of the outcome estimation comparing our approach to existing alternatives on the positivity violations testing dataset ($N = 10,000$) across 50 subsamples

Effect estimation bias reduction with tree depth We see in this case, that having a max depth which is too large leads to an increase in the bias. This might indicate that a hyper parameter tuning procedure would benefit cases where the tree might become too deep and overfit to the training data.

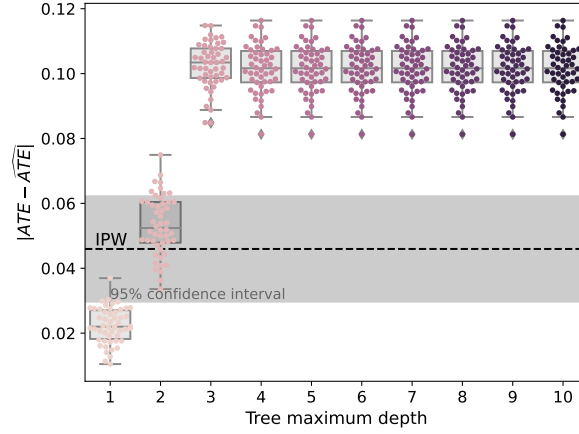


Figure A11: Estimation bias when comparing $BICauseTree(Marginal)$ with varying maximum depth parameters with the average bias of IPW (dotted), on the positivity violations experiment training set ($N = 10,000$) across 50 subsamples.

712 **Treatment allocation bias reduction with tree depth** Figure A12 below shows the weighted
 713 ASMD of all three covariates in $BICauseTree$ models with varying maximum tree depth hyperparam-
 714 eters. We see that the ASMD generally decreases with tree depth, showing that the splitting criteria
 715 we use leads to balanced subpopulations, as expected.

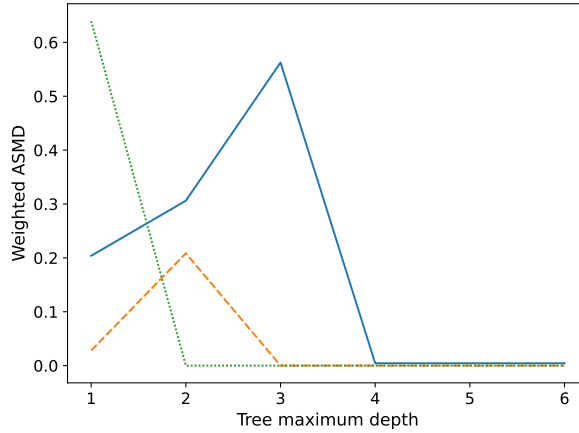


Figure A12: Weighted ASMD for all covariates applying $BICauseTree(Marginal)$ models with varying maximum tree depths on the positivity violations dataset training set ($N = 10,000$) across 50 subsamples

716 A.6 Further experiment results: the twins dataset

717 **Partitioning** The following plot shows the final tree built for the twins dataset. Red nodes indicate
 718 positivity violation population.

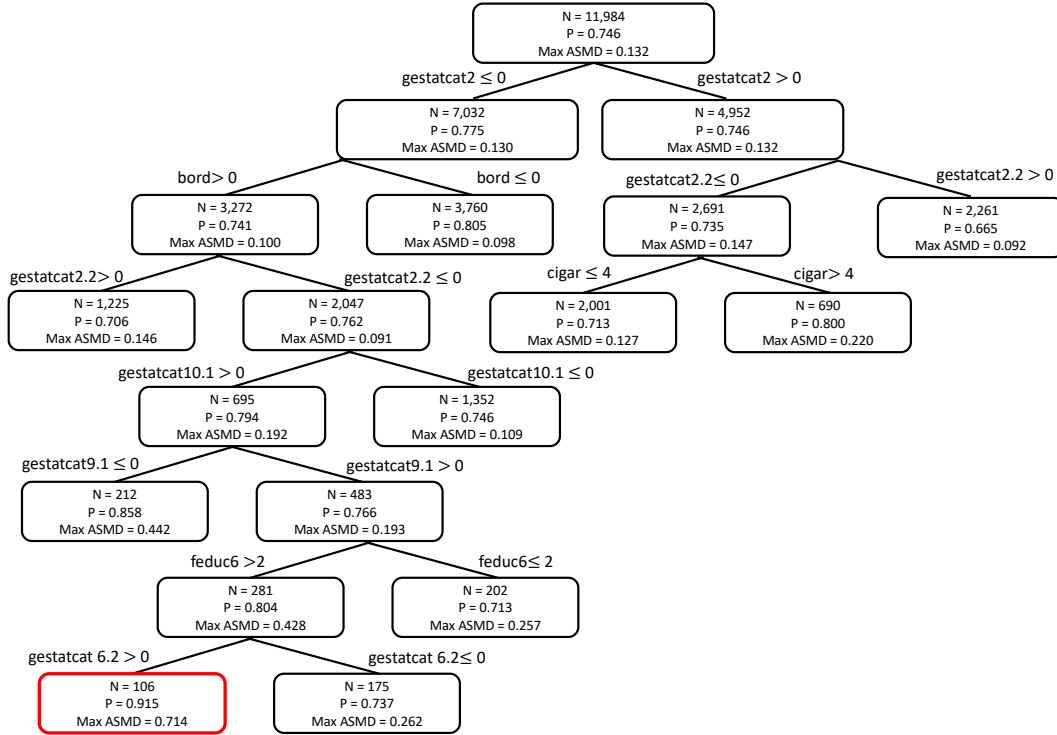


Figure A13: Tree structure after training on the entire twins dataset ($N = 11,984$). Violating leaf nodes are marked in red. Note that this was mistakenly referred to as figure A.4. in the main text, on line 319.

719 **Causal effect estimation** Figure A14 shows a comprehensive comparison of the different mod-
 720 els. Causal tree shows poor performance, whereas causal forest performs comparably to BICause-
 721 Tree(Marginal) and IPW.

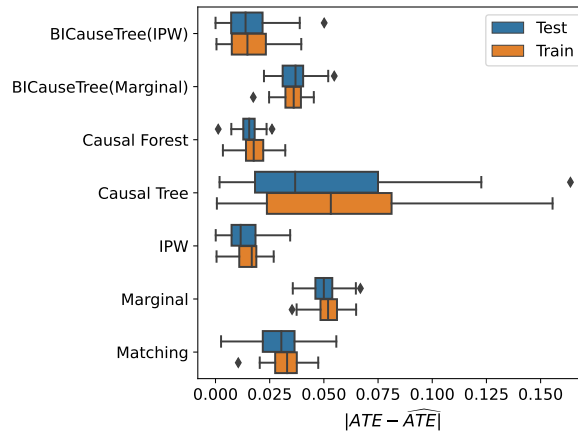


Figure A14: Estimation bias on the twins dataset ($N = 11,984$) across 50 subsamples

722 **Outcome estimation** BICauseTree(Marginal) shows good calibration on outcome estimation. BI-
 723 CauseTree(IPW) seem to have some bias in outcome estimation. This is possibly due to the parametric
 724 nature of our IPW model which uses a Logistic Regression. Matching shows poor outcome estimation
 725 ability.

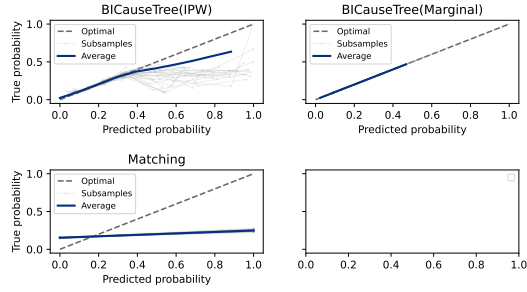


Figure A15: Calibration of the outcome estimation across 50 subsamples, on the twins testing set ($N = 5,992$).

726 **Treatment allocation bias reduction with tree depth** Figure A16 below shows the reduction in
 727 ASMD for the top 10 most imbalanced features in the entire population. It compares BICauseTree
 728 models with varying maximum tree depth hyperparameters. In all 10 covariates, we notice how the
 729 ASMD reduces as the maximum tree depth increases.

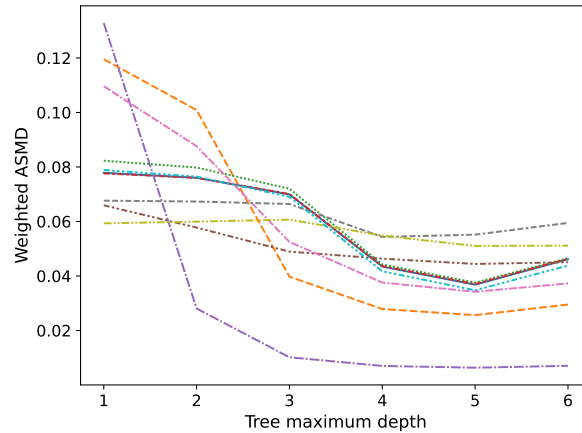


Figure A16: Weighted ASMD for the top 10 covariates with the highest ASMD in the initial population for *BICauseTree(Marginal)* models with varying maximum tree depths on the twins dataset training set ($N = 5,992$) across 50 subsamples

730 A.7 Further experiment results: the ACIC dataset

731 **Partitioning** The following shows the tree built for the ACIC dataset.

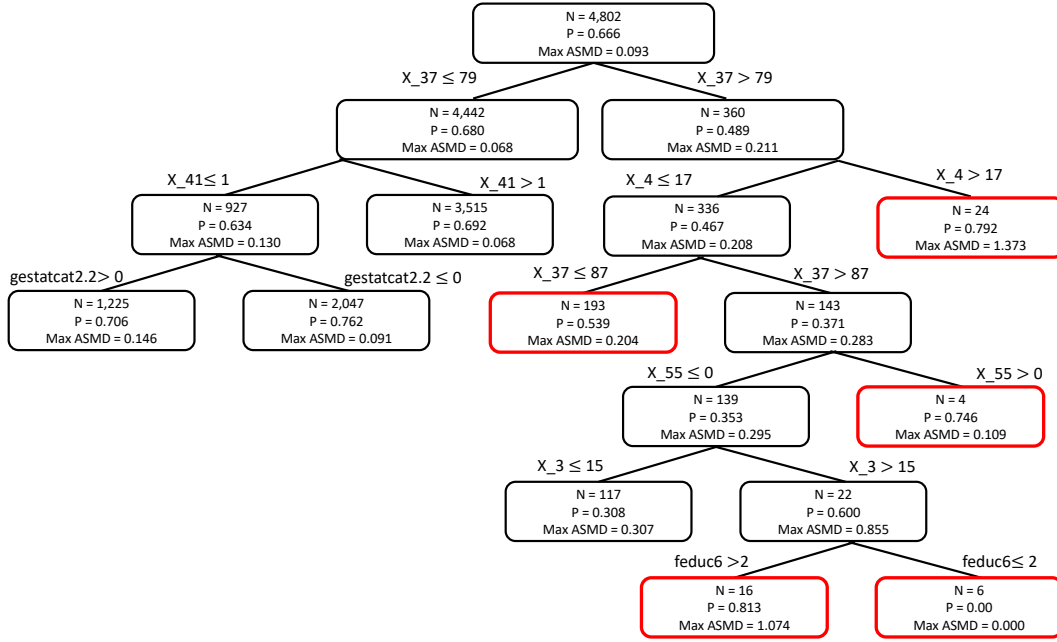


Figure A17: Tree structure after training on the entire ACIC dataset ($N = 4,802$). Violating leaf nodes are marked in red.

732 **Causal effect estimation** Figure A14 shows a comprehensive comparison of the different models.
 733 The performance of both BICauseTree models compares with the one from IPW or Causal Forest.
 734 Causal Tree however shows poor performance. Matching shows high estimation bias compared to all
 735 other models, including the Marginal estimator.

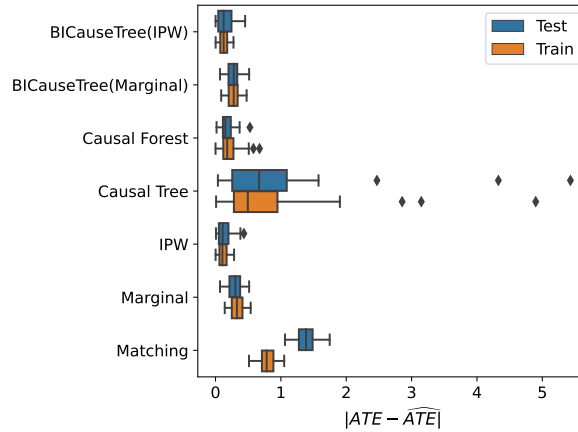


Figure A18: Estimation bias on the ACIC dataset testing set ($N = 960$) across 50 subsamples

736 **Propensity score estimation** We show here the calibration of propensity scores in the ACIC dataset.
 737 BICauseTree preforms less well than IPW, which is more tailored for propensity estimation. As the
 738 outcome is non-binary in ACIC, we are not able to generate a calibration plot comparing the outcome
 739 estimation of BICauseTree with the one from existing approaches.

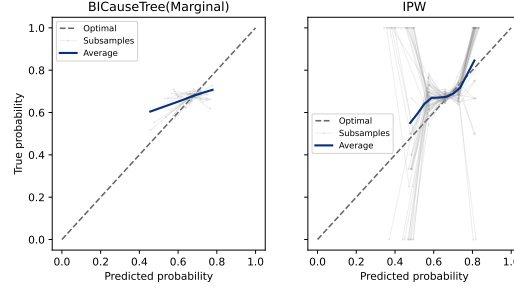


Figure A19: Propensity score calibration comparing our approach to existing alternatives on the ACIC dataset testing set ($N = 960$) across 50 subsamples

Effect estimation bias reduction with tree depth Figure A20 below shows the estimation bias for various maximum tree depth hyperparameters. We notice some overlap to IPW confidence interval, and bias reduces for depths up to max depth 6, then the variance across subsamples starts to increase. This may be due to the limited sample size of this dataset.

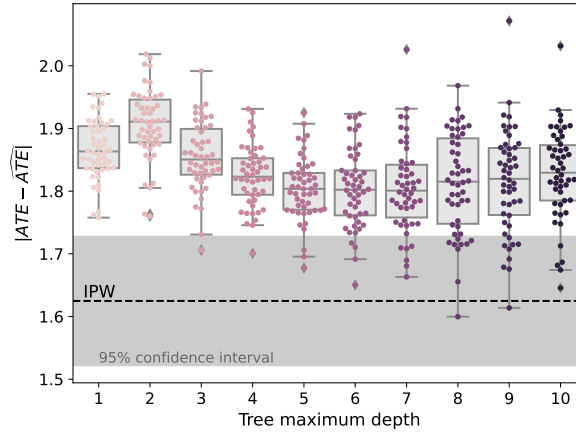


Figure A20: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the ACIC training set ($N = 3,842$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure A21 below shows the reduction in ASMD for the top 10 most imbalanced features in the entire population. It compares *BICauseTree* models with varying maximum tree depth hyperparameters. All 10 covariates show reduction in ASMD with depth.

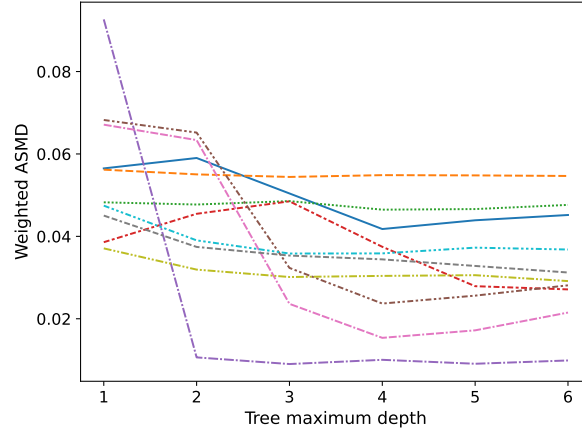


Figure A21: Weighted ASMD across maximum tree depths for the top 10 covariates with the highest ASMD in the initial population, applying *BICauseTree(Marginal)* models with varying maximum tree depths on the ACIC dataset training set ($N = 3,842$) across 50 subsamples

A.8 Implementation

A.8.1 Positivity violations evaluations

We implemented two positivity violations definition procedures: the Crump method and a method we introduced, which we call the *symmetric prevalence* threshold method. The Crump method is a data driven method that defines a threshold for extreme propensity scores (i.e., positivity violations), based on the distribution of propensity scores in the data (see further results in (14)).

The symmetric prevalence procedure generates upper and lower cutoff values that are adjusted for the prevalence. The cutoffs are computed such that if the overall prevalence was 0.5 they would be symmetrical (e.g. for $\alpha = 0.05$ the cutoffs would be 0.05 and 0.95). We may consider this as class reweighting of the propensities, within the entire population.

If we denote the overall prevalence μ , and consider α as the cutoff had the distribution been symmetric (we recommend taking $\alpha = 0.1$), the cutoffs are computed as follows:

$$Upper\ cutoff = \frac{(1 - \alpha) * \mu}{(1 - \alpha) * \mu + \alpha * (1 - \mu)}$$

$$Lower\ cutoff = \frac{\alpha * \mu}{\alpha * \mu + (1 - \alpha) * (1 - \mu)}$$

A.9 Experimental details and computation

For BICauseTree, most hyperparameters were set to their default value. Multiple hypothesis test correction was done following a step-down method using Holm-Bonferroni adjustments (26; 27), with $\alpha = 0.05$. The threshold for weight trimming for positivity violations was computed using the Crump procedure (48; 14) with 10000 segments. Throughout all experiments the minimum treatment group size was set to 2 patients. The maximum depth is the only parameter which varied depending on experiments. It was set to 5 for both synthetic datasets, and 10 for the experiments on the twins dataset. A long-standing practice has been to define any covariate with $ASMD \geq 0.10$ as a potentially problematic confounder (25), so we set this as our default threshold and used it in our experiments.

For IPW(LR), we used a Logistic Regression with a *saga* solver, no penalty and a maximum number of iterations equal to 500. For comparison purposes, BICauseTree(IPW) used similar hyperparameters for its internal IPW outcome model. For IPW(GBT) we used default hyperparameters.

We applied double matching based on the Mahalanobis distance for all datasets but ACIC, on which we used a Euclidean distance to avoid non-invertable matrices caused by the sparsity of non-null column values.

Causal Tree with a single estimator and a subforest size of 1. For Causal Forest, we used 50 estimators and subforest size of 1.

In general BICauseTree compares with IPW, Causal Tree and Causal Forest w.r.t computational efficiency. Matching has higher compute time than all other models.

Experiment	Amount of Compute
Natural experiment dataset	
BICauseTree(Marginal)	286
IPW	134
Matching	882
Causal Tree	183
Causal Forest	390
Positivity violations dataset	
BICauseTree(Marginal)	258
IPW	128
Matching	929
Causal Tree	137
Causal Forest	384
Twins	
BICauseTree(Marginal)	420
BICauseTree(IPW)	567
IPW	329
Matching	2283
Causal Tree	403
Causal Forest	672
ACIC	
BICauseTree(Marginal)	329
BICauseTree(IPW)	376
IPW	239
Matching	1092
Causal Tree	354
Causal Forest	439

Table A1: Total amount of compute in seconds for model fitting across 50 train-test splits, for selected experiments. All experiments were run on a 10 core CPU Apple M1 Pro.

For including Causal Tree into our experiments, we used the code available at <https://github.com/py-why/EconML>. Outcome and propensity models were trained using `sklearn` with default parameters and 500 maximum iterations for Logistic Regression when relevant. Statistical testing was implemented using the `statsmodel` package.

Causal benchmark datasets The **twins dataset** was originally taken from the denominator file at <https://www.nber.org/research/data/linked-birthinfant-death-cohort-data>. However, we use data generated by *Neal et. al* (30), which simulates an observational study from the initial data by selectively hiding one of the twins with a generative approach. The sample size is $N = 11,984$ pairs of twins, with the essential inclusion criterion being that both individuals were born weighing less than 2kg. The mortality rate amongst the lighter twins is 18.9%, and for the heavier 16.4%, for an average treatment effect of -2.5% (which we thus consider as ground truth). A total of 75 covariates were recorded, relating to the parents' socio-demographic features and medical history, the pregnancy and the birth.

The **ACIC dataset** was generated using the *causalib* package at https://github.com/BiomedSciAI/causalib/blob/master/causalib/datasets/data/acic_challenge_2016/README.md. It contains covariates, simulated treatment, and simulated response variables for the causal inference challenge in the 2016 Atlantic Causal Inference Conference (49). For each of 20 conditions, treatment and response data were simulated from real-world data corresponding to 4802 individuals and 58 covariates. After one-hot encoding, a total of 79 covariates was included. More specifically, we used the set of treatment and response variables *zymu 174570858*, with the two expected potential outcomes (μ_0, μ_1).

A.10 Social impact of our work: further details

Recent years have seen a surge in the Explainable AI (XAI) literature, motivated by rising ethical concerns around artificial intelligence. Model scrutiny is particularly relevant in sensitive environments such as healthcare, which require high safety standards considering the major consequences of invalid predictions on individual trajectories (50). Calls for model transparency are further motivated by evidences that supervised machine learning is inclined to reproduce inherent bias and prejudice against discriminated groups (51). Ultimately, XAI aims at building trust in models that ought to be deployed. In recent years, the demand for transparency expanded beyond the research community,

807 notably with incentives from high institutions. The European Union General Data Protection Regula-
808 tion legislation has mandated a “right to explanation” for individual predictions that can “significantly
809 affect” users (52). In other words, algorithmic results should be re-traceable on demand. In parallel,
810 quality control frameworks such as the U.S. Food and Drug Administration guidance have been
811 introduced to ensure the safety of clinical AI (53). By providing interpretable effect estimation, our
812 BICauseTree approach aligns with the mission of increasing transparency for downstream users. As
813 such, it is most likely to have positive social impact. We however caution against relying exclusively
814 on causal inference when data accuracy or volumes are insufficient, as misleading effect estimates
815 may have negative social impact.