The Bone & Joint Journal

Supplementary Material

10.1302/0301-620X.106B7.BJJ-2024-0267

A) Causal Forest Method Details

For each patient $i = 1, \ldots, 1{,}187$ in the dataset, we observe a binary treatment indicator D_i (1 = *cemented hemiarthroplasty,* 0 = *uncemented hemiarthroplasty*), a matrix of baseline characteristics of patient i denoted by X_i that includes the 20 covariates that may act as treatment modifiers (Table i) and a set of outcomes, $Y_{i,j}$, where j indexes the outcomes. The outcomes are listed in Table ii. We consider a generic outcome Y_i for the methods' description. We describe the method using the Neyman-Rubin potential outcomes framework.[1,2]

Theoretically, two potential outcomes are possible for each patient i: $Y_i(0)$ corresponding to the scenario where patient i is assigned to the uncemented hemiarthroplasty group, and $Y_i(1)$ signifying the outcome had patient i been assigned to the cemented group. However, the fundamental problem of causal inference[3] manifests since, at most, one of the two potential outcomes is ever observed for each patient. The observed outcome Y_i can be represented as $Y_i = D_i \times Y_i(1) + (1 - D_i) \times Y_i(0)$, and the effect of the intervention on the outcome for patient i will be $\tau_i = Y_i(1) - Y_i(0)$.

In this study, the estimands of interest can be obtained by aggregating the τi's: the average treatment effect (ATE) quantifies the overall effect on the population, and the conditional average treatment effects (CATE) quantify the average patient-level effect given their baseline characteristics (Xi = x), which can then be aggregated for subgroups of interest.

$$ATE = E(\tau_i) \tag{1}$$

$$CATE(x) = \tau_i(x) = E(\tau_i | X_i = x) \tag{2}$$

When incorporating the covariates X into the model, a reformulation of the observed outcome Y can be expressed as follows:[4]

$$Y_i = \mu_i(X) + D_i \times \tau_i(X) + E \tag{3}$$

Where μi(X) represents the prognostic effect that results from the impact of a subset of covariates X, while the subset of treatment moderators are included in τi(X). If the treatment assignment is assumed to be non-deterministic, the conditional mean ofY will be represented as:[4]

$$E(Y_i|X_i = x) = \mu_i(x) + e_i(x) \times \tau_i(x) = m_i(x) \tag{4}$$

where $e_i(x)$ is the propensity score that is estimated by regressing the treatment on the covariates, and $m_i(x)$ is referred to as the marginal mean.

A.1) Causal forest

To estimate the $CATE(x)$, we apply the causal forest method,[5] which is a generalization of the random forest of Breiman[6] to the estimation of treatment effects. Athey and Imbens[7] modified the classification and regression tree (CART) prediction approach to construct a 'causal tree' which focuses on estimating the expected conditional treatment effects, $\tau_i(x)$, rather than predicting the outcome (Y_i), as is done in a traditional CART. To achieve this, equation (3) is rewritten as:[5]

$$\begin{aligned}(Y_i|X_i = x) &= m_i(x) - m_i(x) + \mu_i(X) + D_i \times \tau_i(X) + E \\ &= m_i(x) + \tau_i(X)(D_i - e_i(x)) + E\end{aligned} \tag{5}$$

This representation enables the estimation of the treatment effects $\hat{\tau}_i(x)$ through a two-step process initiating by regression of outcome and treatment on covariates to obtain estimates of marginal mean $\hat{m}_i(x)$ and the propensity $\hat{e}_i(x)$, respectively. Subsequently, the estimates of interest $\hat{\tau}_i(x)$ are derived by selecting $\hat{\tau}_i(\mathbf{X})$ which minimizes the loss function as defined by Equation (6):[4]

$$\frac{1}{2}[Y_i - \hat{m}_i(x) - \hat{\tau}_i(x)(D_i - \hat{e}_i(x))]^2 \tag{6}$$

This local centring algorithm enhances the model's robustness to potential confounding effects.[8]

Furthermore, an 'honest' estimation is implemented where partitioning and estimating the effects are conducted on distinct subsamples to prevent overfitting and provide correct inference. That is, the splitting criterion of the causal tree aims to minimize the expected mean squared error (EMSE) of the treatment effects, is defined as:[7]

$$-EMSE_T(S^{tr}, N^{est}, T) = \frac{1}{N^{tr}}\sum_{i\in S^{tr}} \hat{\tau}^2\,(X_i|S^{tr}, T)) - (\frac{1}{N^{tr}} + \frac{1}{N^{est}})\sum_{L\in T}(\frac{S^2{}_{S^{tr}_{cemented^{(L)}}}}{p} + \frac{S^2_{S^{tr}_{uncemented^{(L)}}}}{1-p}) \qquad (7)$$

where, S^{train} is the training subsample that is used to construct the tree T, S^{est} is the estimation subsample which is different from the training subsample, N^{est} is the number of patients in the estimation sample, N^{tr} is the number of patients in the training subsample, L is a 'leaf' (i.e. a subgroup defined by the splits) in tree T, $S^2{}_{S^{tr}_{cemented^{(L)}}}$ and $S^2_{S^{tr}_{uncemented^{(L)}}}$ are the within-leaf variances of outcomes for the patients at the two treatment arms, and p is the marginal treatment probability $P(D_i = 1)$ which is constant and does not depend on $\mathbf{X}_i$ in fully randomized experiments such as the WHiTE 5 trial considered here.

This splitting criterion is constructed to prefer leaves exhibiting heterogeneous effects by maximizing the first term of equation (7), and simultaneously, leaves with a good fit by minimizing the within-leaf variance. However, an individual tree can be too noisy. To overcome this, Wager and Athey (2018)[5] proposed the CF which generates an ensemble of B causal trees, each of which produces an estimate $\hat{\tau}_b(X)$, which are then aggregated to obtain a CATE estimate, $\hat{\tau}(X)$. The $\hat{\tau}_i(X)$ estimates are estimated using an adaptive locally weighted estimator[9] such that:

$$\hat{\tau}_i(x) = \frac{\sum_{i=1}^{n} \alpha_i\,(x)(Y_i - \hat{m}^{(-i)}(X_i))(D_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(D_i - \hat{e}^{(-i)}(X_i))^2} \qquad (8)$$

where the superscript $(-i)$ denotes the out-of-bag predictions which are obtained from the subsample of trees where observation i was not used to determine the splits, $\hat{m}(x)$ is the estimated conditional mean outcome $E[Y_i|X_i = x]$ obtained by fitting a regression forest, $\hat{e}(x)$ is the estimated conditional propensity score $P[D_i = 1|X_i = x]$

obtained by fitting a binary regression forest, and $\hat{\alpha}(x)$ is the weight given to observation *i* which measures how often observation *i* is assigned to the same leaf that the point (x) lies within.[9] This method is implemented in the generalized random forest R package *grf*.[10] We estimate CATEs for our pre-specified subgroups by taking the estimated patient-level treatment effects and plugging them into an augmented inverse propensity weighting AIPW estimator[11] of group average treatment effects.[12]

A.2) AIPW estimator

The strength of the AIPW estimator[11] stems from its double robustness property, which means that the estimates of the average treatment effects of the population and the subgroups remain consistent even if one of the propensity or outcome regression forests is mis-specified.[13] Glynn and Quinn[13] provided a theoretical and experimental evidence of its superiority over other estimators such as: regression estimator, inverse propensity weighted (IPW) estimator, and propensity score matching estimator.

In our study, the AIPW scores that are averaged to obtain the ATE and CATE estimates are obtained using the following formula:[14]

$$\hat{\gamma}_i = \hat{m}_i(X_i, 1) - \hat{m}_i(X_i, 0) + \frac{Y_i - \hat{m}_i(X_i, D_i))(D_i - \hat{e}(X_i))}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \quad (9)$$

where $\hat{m}_i(x,d) = \mathrm{E}[Y_i(d) | X_i = x]$ denotes the nonparametric estimate of the conditional mean of the treatment group.

B) Application of CF approaches to estimate group ATEs in the WHiTE 5 Trial

We implement the CF for each outcome using 20,000 trees. This number of trees is large enough to make the perturbation error – which results from fitting different forests – negligible to the variances of the estimated CATEs.[10] All other tuning hyperparameters (sample fraction used to build each tree, number of variables tried for each split, minimum number of individuals in each tree leaf, honesty fraction, and parameters which determine the imbalance of the splits) are determined using cross-validation.

The forests were fitted in two stages.[15] During the first stage, the model is fitted over all covariates. The second stage considers only the most important covariates, i.e. those whose importance exceeds 20% of the average importance (see Figure b.1),

where importance is defined as the simple weighted sum of how many times each covariate was used to determine the sample split at each depth in the forest.[9] Then, we regressed the estimated CATEs on the most important covariates, and obtained the estimates of best linear projection along with coefficient standard errors (see Figure a).

To test for heterogeneity, omnibus heterogeneity tests were performed, and their results are presented in Supplementary Table i. This test yields two parameters: ATE parameter to test the null hypothesis of good calibration of the ATE, where a value of 1 indicates a correct mean forest. The second parameter is the Heterogeneity parameter, also with a value of 1 indicating well calibrated estimates of heterogeneity within the forest. If the Heterogeneity parameter is positive, its associated p-value indicates the strength of evidence supporting the null hypothesis of no heterogeneity.[9] However, the calibration tests indicate the absence of heterogeneity, since the heterogeneity parameter is negative for the six outcomes.

Furthermore, we applied the rank-weighted average treatment effect (RATE) metric proposed by Yadlowsky et al[16] to test to examine the presence of substantial heterogeneity, and to assess the strength of our CATE estimates are at distinguishing subpopulations with different treatment effects. Particularly, we aim to measure the benefit there is to prioritizing cemented therapy provision based on the heterogeneity that is identified by our causal forest. This approach assigns, based on the estimated CATEs, a higher score to patients estimated to benefit more from cemented therapy and a lower score to those with lower benefit compared to uncemented one. The benefit refers to the expected increase in outcomes when providing the cemented therapy to a fraction of the population with the highest prioritization scores as opposed to giving the therapy to a randomly selected fraction of the same size. The figures (see Figure a) depict the target operator characteristic (TOC) curves on the outcomes. These curves chop the population up into groups defined by above mentioned scores, then plot this over all groups where each group is the top q-th fraction of patients with the largest score.

Table i. Calibration tests.

Outcome	ATE parameter (SE)	p-value	Heterogeneity parameter (SE)	p-value
1 mth				
EQ-5D Index	1.01 (0.479)	0.02	-1.61 (-1.76)	0.96
EQ-5D VAS	1.01 (0.756)	0.09	-1.14 (0.828)	0.92
4 mths				
EQ-5D Index	1.01 (0.844)	0.12	-0.74 (0.760)	0.84
EQ-5D VAS	1.04 (1.174)	0.19	-1.69 (0.979)	0.96
12 mths				
EQ-5D Index	1.26 (4.789)	0.40	-1.50 (1.052)	0.92
EQ-5D VAS	0.94 (2.186)	0.33	-18.27 (2.203)	1.00

ATE, average treatment effect; EQ-5D, EuroQol five-dimension health questionnaire; SE, standard error; VAS, visual analogue scale.

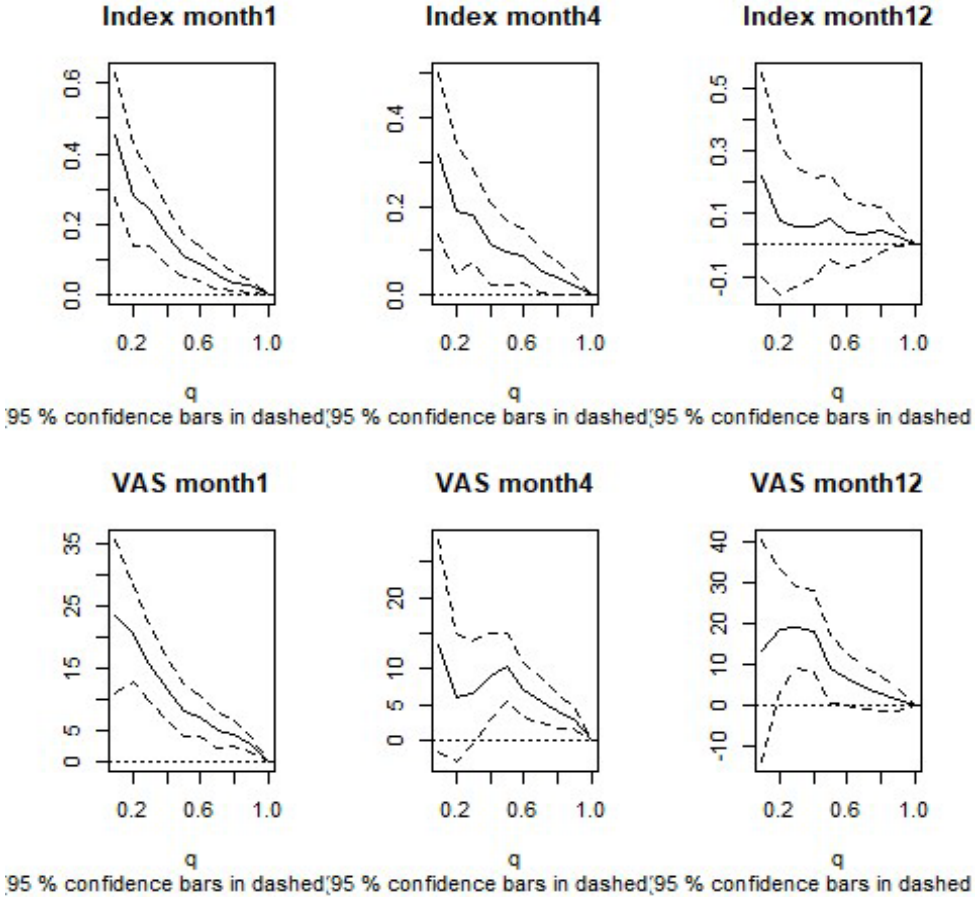

Fig a. Causal forest approaches to estimate group average treatment effects. VAS, visual analogue scale.

Table ii. Outcome characteristics.

Outcome	Overall	Uncemented	Cemented	p-value*
Baseline				
Total, n	956	484	472	
EQindex, Median (IQR)	0.64 (0.38 to 0.80)	0.64 (0.38 to 0.79)	0.65 (0.40 to 0.80)	0.260*
EQ-5D VAS score, Median (IQR)	60.0 (50.0 to 80.0)	65.0 (50.0 to 80.0)	60.0 (50.0 to 80.0)	0.640*
Missing, n (%)	231 (19.5)	111(18.7)	120 (20.3)	
1 mth				
Total, n	827	408	419	
EQindex, Median (IQR)	0.48 (0.08 to 0.66)	0.42 (0.71 to 0.64)	0.53 (0.09 to 0.68)	0.010*
EQ-5D VAS, Median (IQR)	60.0 (45.0 to 75.0)	60.0 (40.0 to 75.0)	60.0 (50.0 to 80.0)	0.060*
Missing, n (%)	360 (30.3)	187 (31.4)	173 (29.2)	
4 mths				
Total, n	579	285	294	
EQindex, Median (IQR)	0.55 (0.12 to 0.70)	0.52 (0.09 to 0.69)	0.57 (0.14 to 0.71)	0.110*
EQ-5D VAS score, Median (IQR)	60.0 (50.0 to 80.0)	60.0 (50.0 to 75.0)	65.0 (50.0 to 80.0)	0.220*
Missing, n (%)	608 (51.2)	310 (52.1)	298 (50.3)	
12 mths				
Total, n	104	117	221	

EQindex, Median (IQR)	0.54 (0.13 to 0.72)	0.52 (0.12 to 0.73)	0.58 (0.16 to 0.71)	0.460*
EQ-5D VAS score, Median (IQR)	60.0 (50.0 to 80.0)	60.0 (50.0 to 80.0)	62.5 (50.0 to 80.0)	0.860*
Missing, n (%)	966 (81.4)	491 (82.5)	475 (80.2)	

EQ-5D, EuroQol five-dimension health questionnaire; IQR, interquartile range; VAS, visual analogue scale.

*Kruskal-Wallis test.

Table iii. Covariates, month 1.

Covariate	Overall	Uncemented	Cemented	p-value
Total, n	827	408	419	
Median age, yrs (IQR)	86.0 (80.0 to 90.0)	86.0 (80.25 to 90.0)	85.0 (80.0 to 90.0)	0.920*
Median EQ-5D (IQR)	0.48 (0.08 to 0.66)	0.42 (0.71 to 0.64)	0.53 (0.09 to 0.68)	0.010*
Median EQ-5D VAS (IQR)	60.0 (45.0 to 75.0)	60.0 (40.0 to 75.0)	60.0 (50.0 to 80.0)	0.060*
Type of consent, n (%)				0.076†
Individual consent	338 (40.9)	182 (44.6)	156 (37.2)	
Proxy consent	408 (49.3)	192 (47.1)	216 (51.6)	
Missing	81 (9.8)	34 (8.3)	47 (11.2)	
Sex, n (%)				0.166†
Male	573 (69.3)	273 (66.9)	300 (71.6)	
Female	254 (30.7)	135 (33.1)	119 (28.4)	
Current smoker, n (%)				0.110†
No	752 (90.9)	377 (92.4)	375 (89.5)	
Yes	62 (7.5)	24 (5.9)	38 (9.1)	
Missing	13 (1.6)	7 (1.7)	6 (1.4)	
Chronic renal failure, n (%)				0.888†
No	762 (92.1)	373 (91.4)	389 (92.8)	
Yes	55 (6.7)	28 (6.9)	27 (6.4)	

Missing	10 (1.2)	7 (1.7)	3 (0.7)	
Diabetes n (%)				0.740†
No	680 (82.2)	336 (82.4)	344 (82.1)	
Yes	135 (16.3)	64 (15.7)	71 (16.9)	
Missing	12 (1.5)	8 (2.0)	4 (1.0)	
Alcohol consumption, n (%)				0.792†
0 to 7 units/wk	735 (88.9)	364 (89.2)	371 (88.5)	
8 to 14 units/wk	39 (4.7)	18 (4.4)	21 (5.0)	
15 to 21 units/wk	17 (2.1)	9 (2.2)	8 (1.9)	
> 21 units/wk	18 (2.2)	7 (1.7)	11 (2.6)	
Missing	18 (2.2)	10 (2.5)	8 (1.9)	
Residence status before injury, n (%)				
Own home/sheltered housing	629 (76.1)	299 (73.3)	330 (78.8)	0.172†
Residential care	93 (11.2)	50 (12.3)	43 (10.3)	
Nursing care	105 (12.7)	59 (14.5)	46 (11.0)	

EQ-5D, EuroQol five-dimension health questionnaire; IQR, interquartile range; VAS, visual analogue scale.

*Kruskal-Wallis test.

†Chi-squared test.

Table iv. Covariates, month 4.

Covariate	Overall	Uncemented	Cemented	p-value
Total, n	579	285	294	
Median age, yrs (IQR)	85.0 (80.0 to 89.0)	85.0 (80.0 to 89.0)	85.0 (79.0 to 89.25)	0.940*
Median EQ-5D (IQR)	0.55 (0.12 to 0.70)	0.52 (0.09 to 0.69)	0.57 (0.14 to 0.71)	0.110*
Median EQ-5D VAS (IQR)	60.0 (50.0 to 80.0)	60.0 (50.0 to 75.0)	65.0 (50.0 to 80.0)	0.220*
Type of consent, n (%)				0.127†
Individual consent	215 (37.1)	115 (40.4)	100 (34.0)	
Proxy consent	292 (50.4)	135 (47.4)	157 (53.4)	
Missing	72 (12.4)	35 (12.3)	37 (12.6)	
Sex, n (%)				0.072†
Male	411 (71.0)	192 (67.4)	219 (74.5)	
Female	168 (29.0)	93 (32.6)	75 (25.5)	
Current smoker, n (%)				0.292†
No	516 (89.1)	257 (90.2)	259 (88.1)	
Yes	49 (8.5)	20 (7.0)	29 (9.9)	
Missing	14 (2.4)	8 (2.8)	6 (2.0)	
Chronic renal failure, n (%)				0.771†
No	529 (91.4)	258 (90.5)	271 (92.2)	
Yes	40 (6.9)	21 (7.4)	19 (6.5)	

Missing	10 (1.7)	6 (2.1)	4 (1.4)	
Diabetes n (%)				0.752†
No	477 (82.4)	232 (81.4)	245 (83.3)	
Yes	92 (15.9)	47 (16.5)	45 (15.3)	
Missing	10 (1.7)	6 (2.1)	4 (1.4)	
Alcohol consumption, n (%)				0.425†
0 to 7 units/wk	506 (87.4)	252 (88.4)	254 (86.4)	
8 to 14 units/wk	34 (5.9)	216 (5.6)	18 (6.1)	
15 to 21 units/wk	9 (1.6)	3 (1.1)	6 (2.0)	
> 21 units/wk	13 (2.2)	4 (1.4)	9 (3.1)	
Missing	17 (2.9)	10 (3.5)	7 (2.4)	
Residence status before injury n (%)				
Own home/sheltered housing	452 (78.1)	217 (76.1)	235 (79.9)	0.486†
Residential care	66 (11.4)	34 (11.9)	32 (10.9)	
Nursing care	61 (10.5)	34 (11.9)	27 (9.2)	

EQ-5D, EuroQol five-dimension health questionnaire; IQR, interquartile range; VAS, visual analogue scale.

*Kruskal-Wallis test.

†Chi-squared test.

Table v. Covariates, month 12.

Covariate	Overall	Uncemented	Cemented	p-value
Total, n	104	117	221	
Median age, yrs (IQR)	85.0 (79.0 to 90.0)	85.0 (79.0 to 89.0)	85.0 (79.0 to 90.0)	0.470*
Median EQ-5D (IQR)	0.54 (0.13 to 0.72)	0.52 (0.12 to 0.73)	0.58 (0.16 to 0.71)	0.460*
Median EQ-5D VAS (IQR)	60.0 (50.0 to 80.0)	60.0 (50.0 to 80.0)	62.5 (50.0 to 80.0)	0.860*
Type of consent, n (%)				0.004†
Individual consent	73 (33.0)	45 (43.3)	28 (23.9)	
Proxy consent	115 (52.0)	45 (43.3)	70 (59.8)	
Missing	33 (14.9)	14 (13.5)	19 (16.2)	
Sex, n (%)				0.389†
Male	156 (70.6)	70 (67.3)	86 (73.5)	
Female	65 (29.4)	34 (32.7)	31 (26.5)	
Current smoker, n (%)				0.074†
No	199 (90.0)	98 (94.2)	101 (86.3)	
Yes	17 (7.7)	4 (3.8)	13 (11.1)	
Missing	5 (2.3)	2 (1.9)	3 (2.6)	
Chronic renal failure, n (%)				0.041†
No	207 (93.7)	94 (90.4)	113 (96.6)	
Yes	11 (5.0)	9 (8.7)	2 (1.7)	

Missing	3 (1.4)	1 (1.0)	2 (1.7)	
Diabetes n (%)				0.913†
No	190 (86.0)	89 (85.6)	101 (86.3)	
Yes	28 (12.7)	14 (13.5)	14 (12.0)	
Missing	3 (1.4)	1 (1.0)	2 (1.7)	
Alcohol consumption, n (%)				0.272†
0 to 7 units/wk	194 (87.8)	95 (91.3)	99 (84.6)	
8 to 14 units/wk	13 (5.9)	4 (3.8)	9 (7.7)	
15 to 21 units/wk	2 (0.9)	0 (0)	2 (1.7)	
> 21 units/wk	6 (2.7)	2 (1.9)	4 (3.4)	
Missing	6 (2.7)	3 (2.9)	3 (2.6)	
Residence status before injury n (%)				0.441†
Own home/sheltered housing	182 (82.4)	84 (80.8)	98 (83.8)	
Residential care	21 (9.5)	9 (8.7)	12 (10.3)	
Nursing care	18 (8.1)	11 (10.6)	7 (6.0)	

EQ-5D, EuroQol five-dimension health questionnaire; IQR, interquartile range; VAS, visual analogue scale.
*Kruskal-Wallis test.
†Chi-squared test.

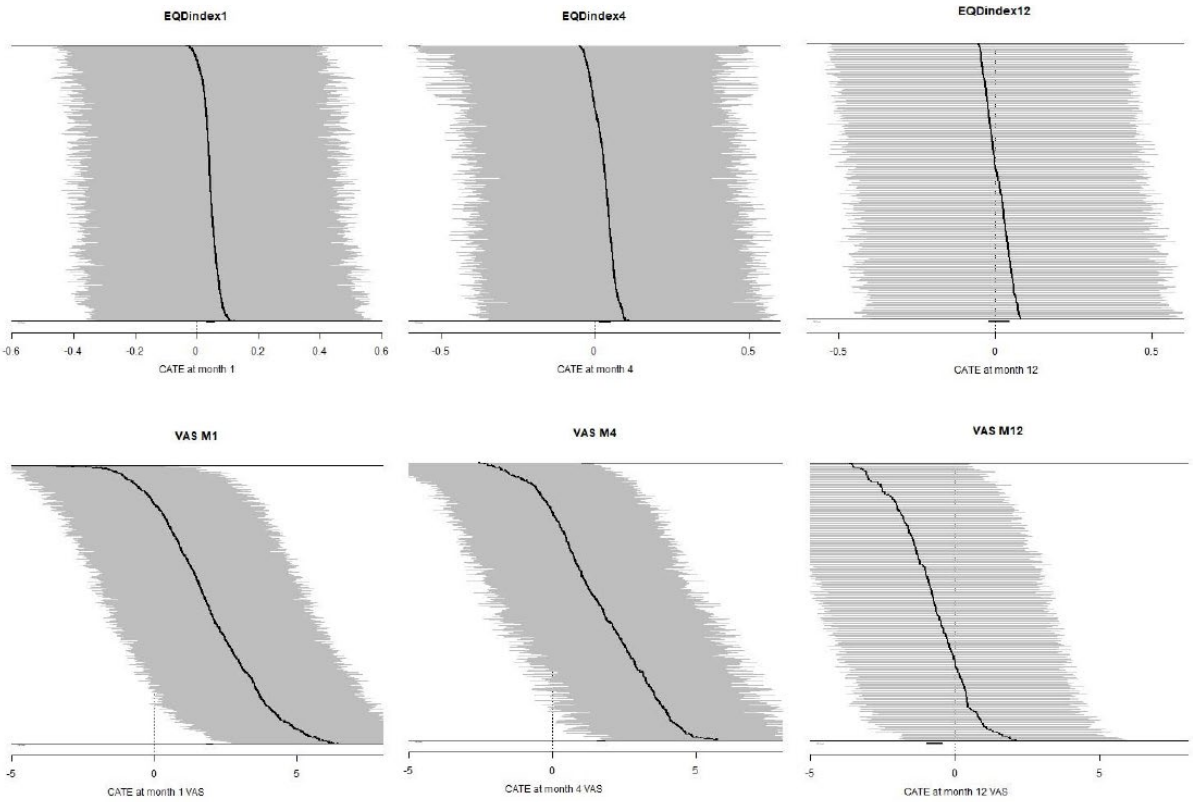

Fig b. Caterpillar plots. This graph shows the individualized effect. CATE, conditional average treatment effect; VAS, visual analogue scale.

References

1. **Rubin DB.** Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688–701.
2. **Splawa-Neyman J, Dabrowska DM, Speed TP.** On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist Sci*. 1990;5(4).
3. **Holland PW.** Statistics and causal inference. *J Am Stat Assoc*. 1986;81(396):945–960.
4. **Dandl S, Hothorn T, Seibold H, Sverdrup E, Wager S, Zeileis A.** What makes forest-based heterogeneous treatment effect estimators work. arXiv. 2022.
5. **Wager S, Athey S.** Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat*. 2018;113(523):1228–1242.
6. **Breiman L.** Random forests. *Mach Learn*. 2001;45(1):5–32.
7. **Athey S, Imbens G.** Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci US*A. 2016;113(27):7353–7360.
8. **Nie X, Wager S.** Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2021;108(2):299–319.
9. **Athey S, Wager S.** Estimating treatment effects with causal forests: an application. *Observational Studies*. 2019;5(2):37–51.
10. **Tibshirani J, Athey S, Stefan Wager RF, Miner L, Wright M.** Grf: generalized random forests. R package version. 2020;1:7–3.
11. **Robins JM, Rotnitzky A, Zhao LP.** Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846–866.
12. **Kreif N, DiazOrdaz K, Moreno-Serra R, Mirelman A, Hidayat T, Suhrcke M.** Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. *Health Serv Outcomes Res Method*. 2022;22(2):192–227.
13. **Glynn AN, Quinn KM.** An introduction to the augmented inverse propensity weighted estimator. *Polit Anal*. 2010;18(1):36–56.
14. **Athey S, Wager S.** Policy learning with observational data. *ECTA*. 2021;89(1):133–161.
15. **Sadique Z, Grieve R, Diaz-Ordaz K, Mouncey P, Lamontagne F, O'Neill S.** A machine-learning approach for estimating subgroup- and individual-level treatment effects: an illustration using the 65 trial. *Med Decis Making*. 2022;42(7):923–936.
16. **Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S.** Evaluating treatment prioritization rules via rank-weighted average treatment effects. arXiv. 2021.