A Semi-Parametric BART Model for Causal Inference

Interpretable treatment effect heterogeneity

Yusufhan Balci Mark van de Wiel Jeroen Hoogland

BAYES2026, Leiden, The Netherlands
22 October 2026

Amsterdam UMC / VU Amsterdam

Conflict of Interest & Disclosures

Conflict of Interest:

- I have no current or past relationships with commercial entities.

Data Use Disclosure:

> "This study, carried out under YODA Project # 2025-0684, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The original proposal can be found: `https://yoda.yale.edu/data-request/2025-0684/`"

Roadmap

1. **Clinical Question:** Differences in Expected Benefit
2. **Methodology:** Semi-Parametric BART
3. **Evaluation:** Accuracy and Uncertainty
4. **Clinical Application:** The CANVAS Program

Treatment Effect Heterogeneity in Drug Development

Clinical questions

- *Which patient groups* are expected to benefit from treatment?
- *Which baseline characteristics* modify expected benefit?
- *How certain* are the estimated benefits?

Our target is the **conditional average treatment effect (CATE)**:

$$\tau(x) = \mathrm{E}\{Y(1) - Y(0) \mid X = x\}.$$

The methodological challenge

- Separate subgroup analyses face low power and multiplicity (Kent et al.2020).
- Flexible CATE models, such as BCF, offer limited direct inference on effect modifiers (Hahn et al.2020).

The Proposed Solution: A Semi-Parametric BART Model

Goal: Flexible prognosis and an interpretable treatment contrast.

Semi-Parametric Causal Model

$$Y_i = \underbrace{\mu(x_i, \hat{\pi}(x_i))}_{\text{Non-parametric Baseline}} + \underbrace{z_i\,\tau(x_i)}_{\text{Parametric CATE}} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- μ: BART with the fitted propensity score $\hat{\pi}(x)$ as an additional predictor.
- τ: main effects and all pairwise interactions,

$$\tau(x) = \alpha + \sum_{j=1}^{p} \beta_j x_j + \sum_{j<k} \beta_{jk} x_j x_k.$$

Navigating High Dimensions: The Horseshoe Prior

The challenge: p main effects and $p(p-1)/2$ pairwise interactions.

Global–Local Shrinkage

$$\begin{aligned}
\beta_j \mid \lambda_j, \sigma &\sim \mathcal{N}(0, \sigma^2 \lambda_j^2), \\
\beta_{jk} \mid \lambda_{jk}, \lambda_{\text{glob}}, \sigma &\sim \mathcal{N}(0, \sigma^2 \lambda_{jk}^2 \lambda_{\text{glob}}^2), \\
\lambda_j, \lambda_{jk}, \lambda_{\text{glob}} &\sim \mathsf{C}^+(0, 1).
\end{aligned}$$

- Local scales allow strong signals to escape shrinkage.
- The additional global scale regularizes **interactions only**.
- Continuous shrinkage concentrates noise coefficients **near zero**.

Simulation Study: Efficiency vs. Misspecification

Two prognoses, four CATE structures, $N = 250$–3000, **200 replicates**.

Quoted cells: $N = 3000$, non-linear prognosis. Values: SP-BART vs. BCF.

Linear Interaction
$\tau(x) = 1 + 2x_2x_4$

- CATE RMSE: **0.123** vs. 0.467.
- CATE coverage: **98%** vs. 69%.

Heavy Saturation
$\tau(x) = 0.5 + 4/(1 + e^{-2.5x_1})$

- BCF better estimates effect magnitudes.
- RMSE: 0.445 vs. **0.243**.
- CATE coverage: 40% vs. **88%**.

Heavy-saturation ranking: Spearman $\rho = 0.997$ (SP-BART) vs. 0.979 (BCF), despite biased magnitudes.

IHDP: A Semi-Synthetic Test of CATE Accuracy

Real covariates from 747 children, 100 synthetic response surfaces, 10-fold prediction.

Method	Mean CATE RMSE	Median RMSE	Spearman ρ
SP-BART	**1.839**	0.976	**0.882**
BCF	2.016	**0.919**	**0.882**
DR-learner	2.244	1.488	0.751
S-learner	3.108	1.542	0.818
R-learner	4.206	1.910	0.732

- BCF has lower RMSE in **66 of 100** response surfaces.
- SP-BART wins 1/25 in the weakest signal quartile and 20/25 in the strongest.

IHDP: Uncertainty and Coverage-Matched Inference

Raw 95% credible intervals under-cover for both methods.

	SP-BART	BCF
Raw coverage	89.7%	81.6%
Raw intervals excluding zero	89.7%	92.3%
Interval expansion factor	1.269	1.783
Coverage after calibration	95.0%	94.9%
Calibrated intervals excluding zero	86.4%	85.5%

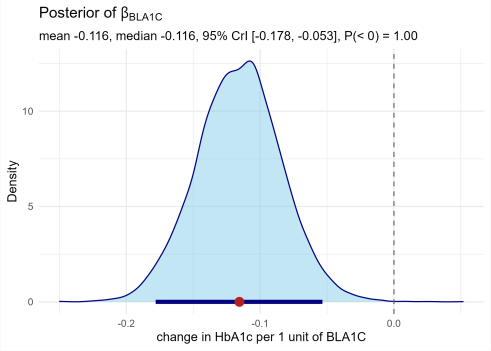
At comparable coverage, exclusion rates are similar.

Calibration uses the other 99 synthetic response surfaces. This does not establish coverage in new patient samples.

Clinical Application: The CANVAS Program

Canagliflozin vs. placebo: change in HbA1c from baseline to 52 weeks.
Training: CANVAS-R ($N = 5018$). Validation: CANVAS ($N = 3004$).
More negative change means greater HbA1c reduction.

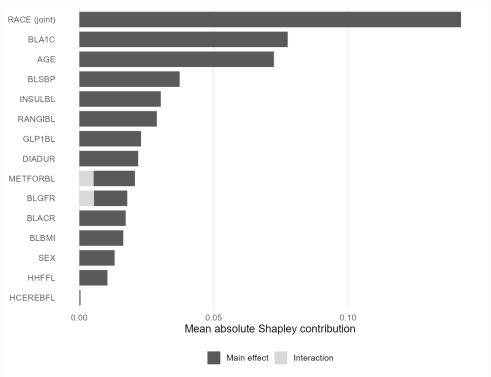
- **Outcome RMSE:** 0.910 vs. 0.950 (SP-BART vs. BCF).
- **CATE agreement:** $r = 0.840$, $\rho = 0.843$.
- **Baseline HbA1c:** mean -0.116, 95% CrI $[-0.178, -0.053]$.
- **Targeting:** uplift near random. True CATEs are unobserved.



Interpreting the CATE via Shapley Values

Explaining fitted CATEs with analytical Shapley values (van de Wiel et al.2024).

- Main effects and interactions contribute to each covariate.
- **Largest contributions:** grouped RACE, HbA1c, and age.
- Simultaneous 95% intervals for their full-CATE contrasts exclude zero.



The RACE association is exploratory and coding-sensitive. These contrasts do not identify effects of intervening on covariates.

Implications for Clinical Treatment Effect Modelling

Key Takeaways

- **Structured treatment effects:** gains in estimation when main effects and interactions provide a useful approximation.
- **Misspecification:** effect magnitudes and coverage can suffer despite accurate ranking.
- **Clinical use:** exploratory modifier assessment with direct inference and analytical Shapley contributions.

CANVAS: limited targeting gains. Treatment selection needs further validation.

Thank you! Questions?

h.y.h.balci@amsterdamumc.nl

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

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

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

References ii

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of internal medicine*, 172(1):35–45.

van de Wiel, M. A., Amestoy, M., and Hoogland, J. (2024). Linked shrinkage to improve estimation of interaction effects in regression models. *Epidemiologic Methods*, 13(1):20230039.

Zeldow, B., Lo Re, III, V., and Roy, J. (2019). A semiparametric modeling approach using Bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. *The annals of applied statistics*, 13(3):1989–2010.

Appendix: Causal Identification and Model Scope

Identification of conditional mean potential outcomes requires:

- **Consistency:** the observed outcome equals the potential outcome under the treatment received.
- **Conditional exchangeability:** $\{Y(0), Y(1)\} \perp Z \mid X$.
- **Overlap:** $0 < P(Z = 1 \mid X = x) < 1$ on the target support.

Conditional Mean Potential Outcomes

$$\mathrm{E}\{Y(0) \mid x\} = \mu(x, \hat{\pi}(x)), \qquad \mathrm{E}\{Y(1) \mid x\} = \mu(x, \hat{\pi}(x)) + \tau(x).$$

Randomization supports exchangeability. Observational analyses require adequate measured confounders.

Appendix: Prior and Posterior Computation

$$\beta_j \mid \lambda_j, \sigma \sim \mathcal{N}(0, \sigma^2 \lambda_j^2), \qquad \beta_{jk} \mid \lambda_{jk}, \lambda_{\text{glob}}, \sigma \sim \mathcal{N}(0, \sigma^2 \lambda_{jk}^2 \lambda_{\text{glob}}^2),$$
$$\lambda_j, \lambda_{jk}, \lambda_{\text{glob}} \sim \mathsf{C}^+(0, 1), \qquad \alpha \sim \mathcal{N}(0, 10^2),$$
$$\sigma^2 \sim \mathsf{InvGamma}(1, 0.001).$$

Bayesian backfitting

1. Update BART using the residual $Y - z\tau(x)$.
2. Update α and β using the residual $Y - \mu(x, \hat{\pi})$.
3. Update local/global shrinkage scales and the residual variance.

The variance update includes the coefficient-prior contribution. BCF uses a BART prior for both μ and τ.

Appendix: Simulation Results at $N = 3000$

		CATE RMSE		Coverage (%)	
CATE structure	Prognosis	SP-BART	BCF	SP-BART	BCF
Linear interaction	Linear	0.143	0.452	97.0	70.0
Linear interaction	Non-linear	0.123	0.467	98.0	69.0
Mild saturation	Linear	0.134	0.217	97.0	92.0
Mild saturation	Non-linear	0.113	0.192	98.0	94.0
Heavy saturation	Linear	0.452	0.263	43.0	86.0
Heavy saturation	Non-linear	0.445	0.243	40.0	88.0
Homogeneous	Linear	0.131	0.159	97.0	96.6
Homogeneous	Non-linear	0.109	0.151	98.6	98.4

200 replicates per cell. Nominal CATE coverage: 95%. Homogeneous: $\tau(x) = 3$.
Saturation: $\tau(x) = 0.5 + 4/(1 + e^{-ax_1})$, with $a = 0.4$ (mild) or 2.5 (heavy).

Appendix: IHDP Accuracy by Signal Strength

Quartiles of the standard deviation of the true CATE

Quartile	Range of sd(τ)	Mean Δ	Median Δ	SP-BART wins
1 (weakest)	0.03–1.08	0.144	0.138	1/25
2	1.15–2.54	0.098	0.051	5/25
3	2.57–5.55	0.050	0.061	8/25
4 (strongest)	5.67–43.22	−1.001	−0.554	20/25

$\Delta = \text{RMSE}_{\text{SP-BART}} - \text{RMSE}_{\text{BCF}}$. Negative values favour SP-BART.

- Each quartile contains 25 response surfaces on the same children.
- Signal strength, scale and shape vary together.

Appendix: Coverage-Matched IHDP Intervals

$$I_i(c) = [m_i + c\{q_{i,0.025} - m_i\},\ m_i + c\{q_{i,0.975} - m_i\}]\,.$$

m_i: posterior mean. $q_{i,\gamma}$: posterior quantile.

For each held-out surface, choose the smallest c attaining 95% empirical coverage on the other 99.

	c	Coverage	Excludes zero	Mean width
SP-BART	1.269	95.0%	86.4%	5.93
BCF	1.783	94.9%	85.5%	6.00

Paired exclusion-rate difference: 0.0095 (95% CI $[0.0025, 0.0164]$).

Empirical calibration across fixed-patient response surfaces does not establish coverage in new patients or a 95% posterior probability interpretation.

Appendix: CANVAS Out-of-Sample Performance

Training: CANVAS-R. Validation: CANVAS ($N = 3004$).

		Outcome RMSE		CATE agreement		AUUC	
N	Draws	BCF	SP-BART	r	ρ	BCF	SP-BART
500	20	1.009	0.949	0.727	0.773	1133	1137
1000	20	0.974	0.926	0.741	0.768	1168	1177
1500	20	0.952	0.919	0.769	0.788	1172	1192
2000	20	0.979	0.917	0.791	0.806	1182	1187
3000	20	0.970	0.913	0.834	0.842	1188	1206
4000	20	0.944	0.910	0.832	0.833	1195	1216
5018	1	0.950	0.910	0.840	0.843	1202	1224

Means over 20 training draws (one full-data fit). True clinical CATEs are unobserved.
CATE agreement compares models. Uplift is close to random targeting.

Appendix: Analytical Shapley Decomposition

For centered covariates $u = x - \mathrm{E}(X)$ and coefficients b expressed in that basis:

$$\phi_j(u) = \underbrace{b_j u_j}_{\text{Main effect}} + \underbrace{\frac{1}{2} \sum_{k \neq j} b_{jk} \left\{ u_j u_k - \mathrm{E}(u_j u_k) \right\}}_{\text{Interaction contribution}}.$$

$$\sum_j \phi_j(u) = \tau(x) - \mathrm{E}\{\tau(X)\}.$$

- Categorical variables: sum signed column contributions within each person before taking absolute values.
- Bars stack mean absolute main and interaction contributions.
- Components can cancel, so this can exceed the mean absolute total.

Appendix: Simultaneous CATE Contrasts in CANVAS

Contrasts of the full fitted CATE, including interactions

- Continuous variables: 90th vs. 10th percentile.
- Binary variables: the two levels. RACE: all category pairs.
- Average each contrast over the empirical values of the other covariates.

For posterior draws $d_\ell^{(m)}$, posterior mean $\bar{d}_\ell$ and SD s_ℓ:

$$k = Q_{0.95}\left[\max_\ell \left|\frac{d_\ell^{(m)} - \bar{d}_\ell}{s_\ell}\right|\right], \qquad I_\ell = \bar{d}_\ell \pm k s_\ell.$$

A covariate is flagged if at least one of its simultaneous intervals excludes zero.

These are model-conditional modifier contrasts. They do not identify causal effects of intervening on the covariates. The RACE finding remains exploratory and coding-sensitive.