# Bayesian Fusion Forests

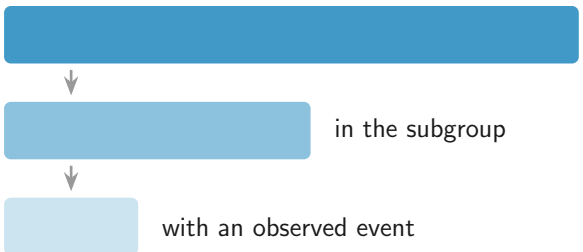
Heterogeneous treatment effects on survival from randomised and real-world data

---

Tijn Jacobs
joint work with Stéphanie L. van der Pas and Wessel N. van Wieringen

VU VRIJE UNIVERSITEIT AMSTERDAM

# Estimating heterogeneous treatment effects is hard



# Randomised and real-world data

**Randomised trial (RCT)**

- ✓ no confounding
- × underpowered
- × short follow-up

**Real-world data (RWD)**

- × unmeasured confounding
- ✓ large
- ✓ long follow-up

# Randomised and real-world data

**Randomised trial (RCT)**

- ✓ no confounding
- × underpowered
- × short follow-up

**Real-world data (RWD)**

- × unmeasured confounding
- ✓ large
- ✓ long follow-up

Challenges for data fusion:

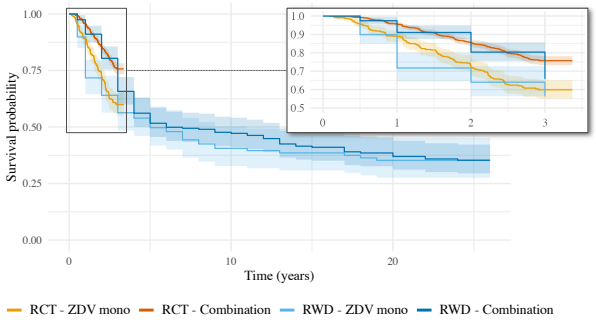
difference in baseline prognosis, measurement protocol and censoring.

# Motivating example: antiretroviral therapy for HIV

Combination therapy vs.
zidovudine (ZDV) therapy.

**ACTG 175** — RCT
$n = 1771$, randomised, $\sim$3 years,
right-censored

**MACS** — RWD
$n = 373$, observational, $\sim$25 years,
**interval-censored**



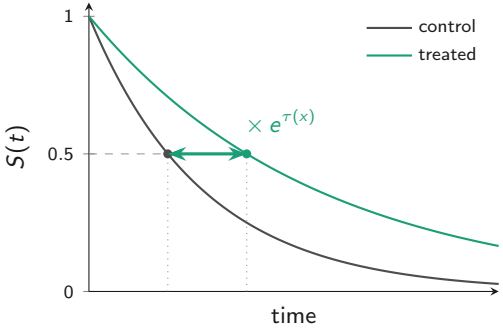
# Estimand: the acceleration factor

Notation:

- $S$: source, RCT (1) or RWD (0)
- $A \in \{0, 1\}$: treatment
- $X$: covariates
- $T(a)$: potential survival time

Treatment effect on the log-time scale:

$$\tau(x) \;=\; \mathbb{E}\big[\log T(1) - \log T(0) \mid X = x\big].$$




$e^{\tau(x)} = 2 \Rightarrow$ median survival doubles.

# Assumptions

We assume consistency, positivity and conditionally independent censoring.

**Unconfoundedness in the RCT.** $T(a) \perp\!\!\!\perp A \mid X, S = 1$.

**Transportability.** The treatment effect is the same in both sources:

$$\mathbb{E}\big[\log T(1) - \log T(0) \mid X = x, S = 1\big] \;=\; \mathbb{E}\big[\log T(1) - \log T(0) \mid X = x, S = 0\big] \;=\; \tau(x).$$

# The confounding function

In the RWD, the observed contrast differs from the treatment effect by the **confounding function**:

$$c(x) = \underbrace{\mathbb{E}[\log T \mid x, A = 1, S = 0] - \mathbb{E}[\log T \mid x, A = 0, S = 0]}_{\text{observed RWD contrast}} - \tau(x)$$

# The confounding function

In the RWD, the observed contrast differs from the treatment effect by the **confounding function**:

$$c(x) \;=\; \underbrace{\mathbb{E}[\log T \mid x, A=1, S=0] - \mathbb{E}[\log T \mid x, A=0, S=0]}_{\text{observed RWD contrast}} \;-\; \tau(x)$$

**Proposition.** The conditional log survival time decomposes as:

$$\mathbb{E}[\log T \mid A,\, X,\, S] \;=\; m_0(X, S) \;+\; \tau(X)\, A \;+\; (1 - S)\, A\, c(X),$$

where $m_0(X, S) := \mathbb{E}[\log T \mid A = 0,\, X,\, S]$ is the baseline log survival time.

# The Bayesian fusion forest

We split the baseline into a shared part and a source deviation to allow for *source-specific heterogeneity*:

$$m_0(X, S) \;=\; m_0^{\mathrm{sh}}(X) \;+\; (1 - S)\, d(X).$$

# The Bayesian fusion forest

We split the baseline into a shared part and a source deviation to allow for *source-specific heterogeneity*:

$$m_0(X, S) \;=\; m_0^{\mathrm{sh}}(X) \;+\; (1 - S)\, d(X).$$

We model:

$$\log T \;=\; \underbrace{m_0^{\mathrm{sh}}(X)}_{\text{shared baseline}} \;+ (1 - S) \underbrace{d(X)}_{\text{source deviation}} \;+ A \underbrace{\tau(X)}_{\text{treatment effect}} \;+ (1 - S)\, A \underbrace{c(X)}_{\text{confounding}} \;+ \varepsilon,$$

- Each function receives its own BART prior, with regularisation tuned to its causal role.
- The errors $\varepsilon$ follow a hierarchical Dirichlet process mixture with source-specific weights.

# Simulation study

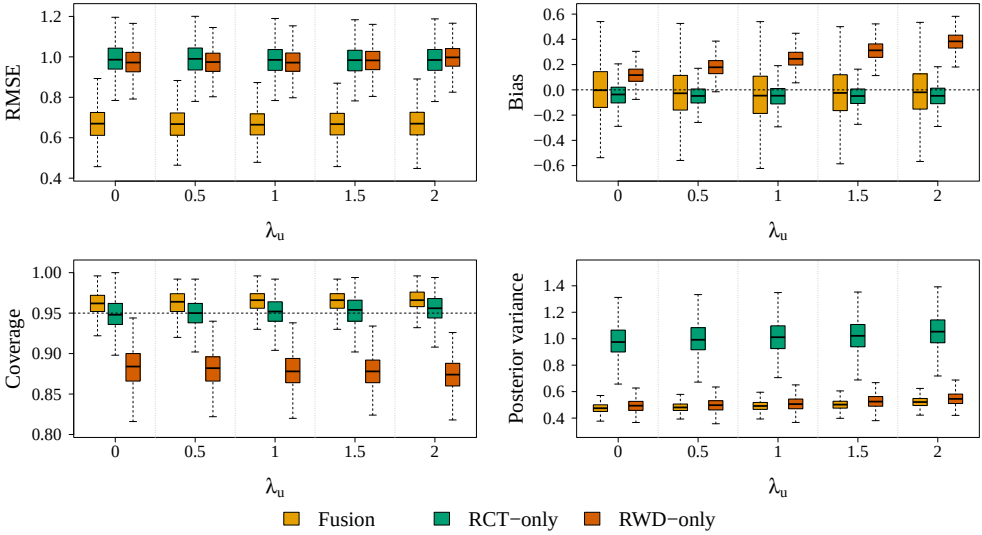
RCT ($n = 150$, right-censored) and RWD ($n = 350$, interval-censored), $p = 10$ covariates.

We vary the unmeasured confounding in the RWD, and compare against RCT-only and RWD-only analyses.

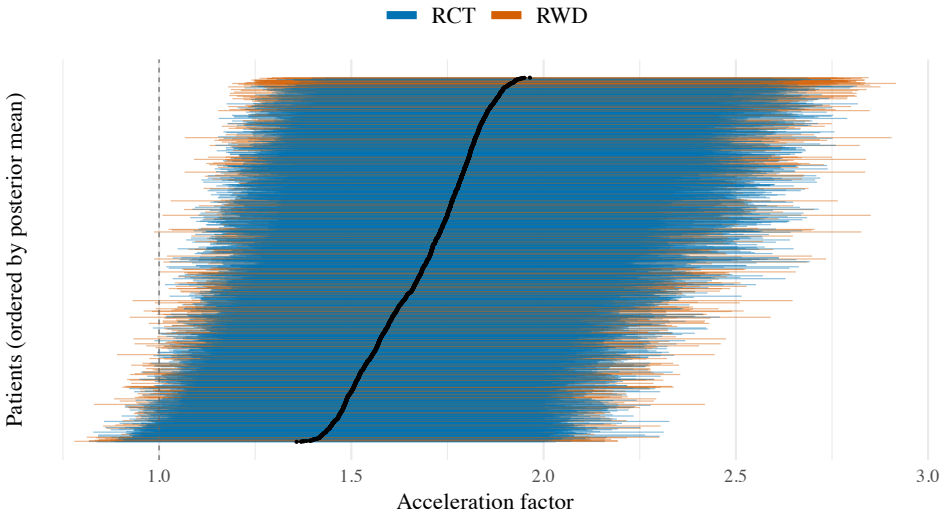
Do we **gain** over the RCT alone?

Do we stay **unbiased** when the RWD is confounded?

# Simulation study



# Posterior of the heterogeneous treatment effects

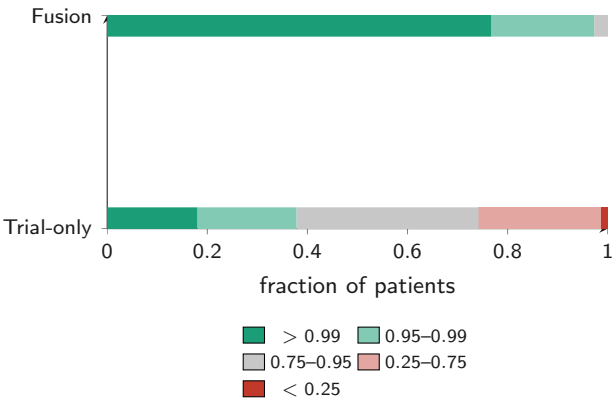


# Posterior probability of benefit

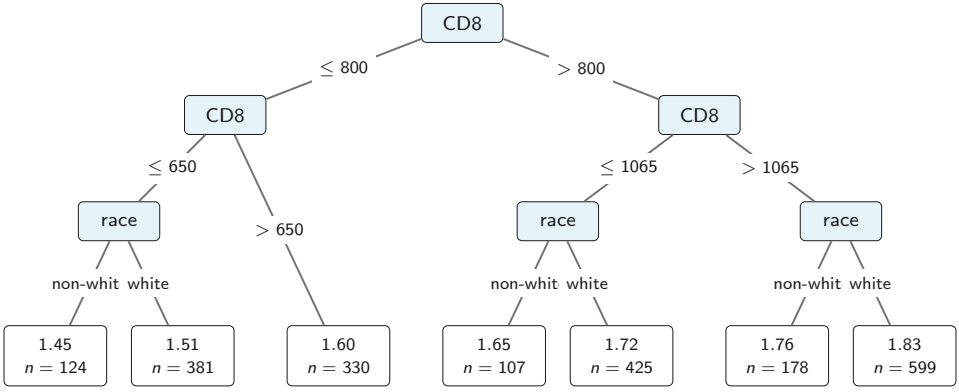
Posterior probability of benefit, $\mathbb{P}(\tau(x_i) > 0 \mid \text{data})$, over the *trial* population.

The trial alone leaves a quarter of patients inconclusive.

The fusion finds benefit for nearly everyone, with a slightly *smaller* average effect.



# Who benefits most?




Posterior projection of $\tau$ onto a single decision tree.

# Take-aways

- **No unconfoundedness in the RWD.** A confounding function absorbs the bias; the treatment effect transports across sources.
- **A separate prior per causal component.** Each forest is regularised for its role.
- **The payoff.** Unbiased under confounding, less posterior variance than an RCT-only analysis, and a benefit for nearly every HIV patient where the RCT alone was inconclusive.

# Thanks for your attention!



**Paper**

Preprint on arXiv.




**R package `FusionForests`**

on GitHub and CRAN.

**Stay up to date**

`tijn-jacobs.github.io`

This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe program under Grant agreement No. 101074802. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

# Assumptions

1. **Consistency.** $A = a \Rightarrow T = T(a), \mathcal{C} = \mathcal{C}(a)$.

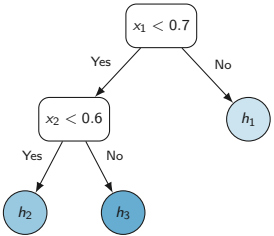
2. **Unconfoundedness of the RCT.** $T(a) \perp\!\!\!\perp A \mid X, S = 1$.

3. **Positivity.** $0 < \mathbb{P}(A = 1 \mid X = x, S = s) < 1$.

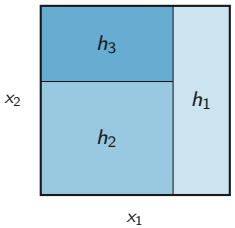
4. **Cross-source transportability.**
   $\mathbb{E}[\log T(1) - \log T(0) \mid X, S = 1] = \mathbb{E}[\log T(1) - \log T(0) \mid X, S = 0]$.

5. **Conditional independent censoring.** $T(a) \perp\!\!\!\perp \mathcal{C}(a) \mid X, A, S$.

# BART in one slide



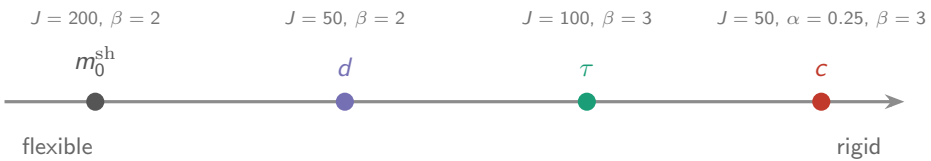



$$f(x) = \sum_{j=1}^{J} g(x; \mathcal{T}_j, \mathcal{H}_j)$$

A sum of $J$ shallow trees. Depth-penalised prior $\alpha(1+d)^{-\beta}$; Gaussian step heights of scale $k$.

# A ladder of regularisation




- $m_0^{\mathrm{sh}}$ enters every observation $\Rightarrow$ let it be complex.
- $d$ is mean-zero $\Rightarrow$ the prior *centres on equality* of the two baselines.
- $\tau$ shrunk towards a homogeneous effect.
- $c$ shrunk towards a constant bias: it comes from what we did not measure.

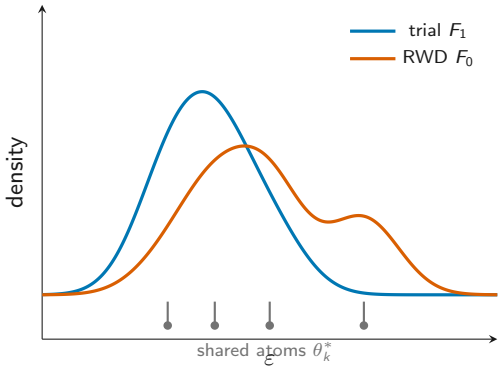
# Error distribution: hierarchical Dirichlet process mixture

Protocol, follow-up quality and extreme outcomes differ between sources.

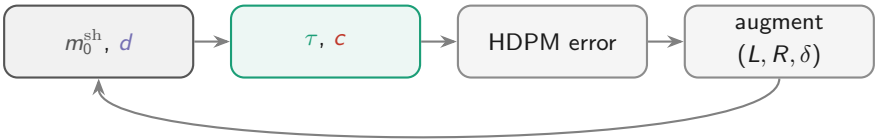
Hierarchical Dirichlet process mixture:

$$G^* \sim \mathrm{DP}(\gamma, H), \quad G_s^* \sim \mathrm{DP}(M_s, G^*).$$

- Shared mixture components.
- Source-specific weights.
- Centred, so $\mathbb{E}[\varepsilon \mid S] = 0$.



# Posterior computation



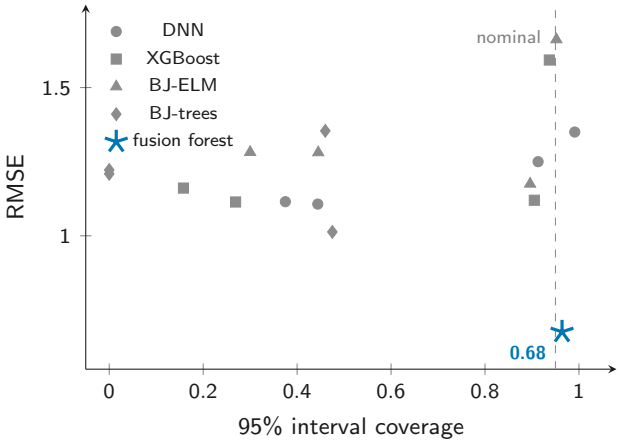

- Blocked Gibbs: each forest is updated on its partial residual, a standard BART step.
- Censored $T_i$ imputed as latent variables (Tanner & Wong, 1987).
- Average effects by hierarchical Bayesian bootstrap over the target covariate distribution.

# Comparison with flexible machine learning

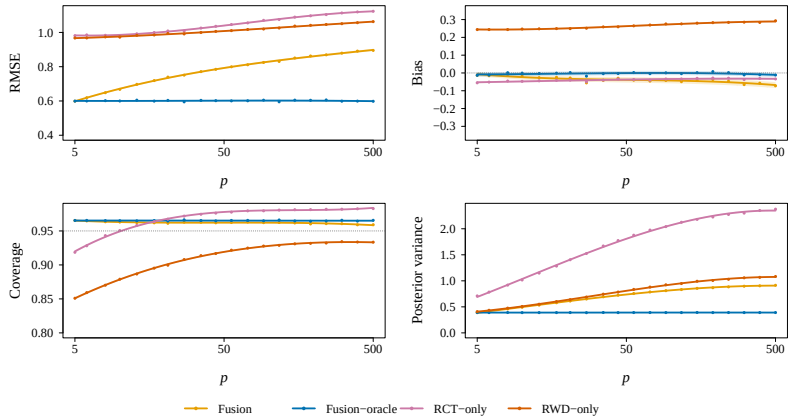
Four predictors $\times$ {S, T} $\times$ {trial, pool}.

None targets a causal contrast; none can separate $\tau$ from $c$.

Only the fusion forest lands in the good corner: lowest RMSE *and* nominal coverage.

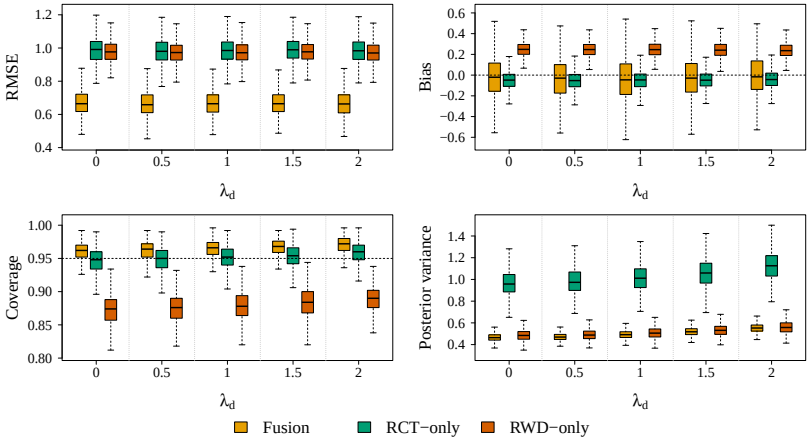


# Varying the covariate dimension $p$



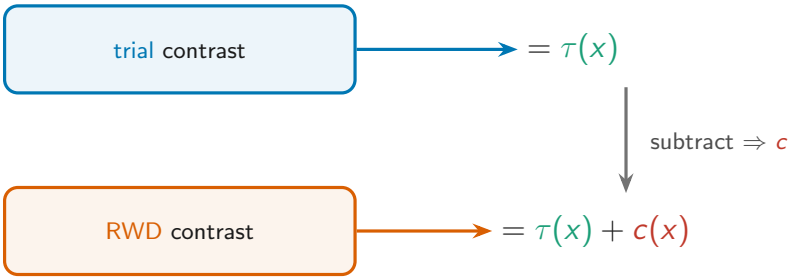

$p : 5 \to 500$, five active. RMSE ratio to RCT-only rises from 0.68 to 0.81.

# Varying the between-source heterogeneity $\lambda_d$




All three estimators are flat in $\lambda_d$: the deviation forest absorbs the gap.

# Identification of the causal effect




The RWD alone identifies only the sum.

The non-identifiability of the causal effect from the RWD is structural and cannot be resolved by collecting more data.

# Full model

$$\log T_i = m_0^{\text{sh}}(X_i) + (1 - S_i)\, d(X_i) + A_i\, \tau(X_i) + (1 - S_i) A_i\, c(X_i) + \varepsilon_i,$$

$$m_0^{\text{sh}} \sim \text{BART}(200,\ k_{\text{sh}},\ 0.95,\ 2),$$

$$d \sim \text{BART}(50,\ k_d,\ 0.95,\ 2),$$

$$\tau \sim \text{BART}(100,\ k_\tau,\ 0.95,\ 3),$$

$$c \sim \text{BART}(50,\ k_c,\ 0.25,\ 3),$$

$$\varepsilon_i \mid S_i = s \sim \int \sigma_s^{-1} \phi\big((w - \theta)/\sigma_s\big)\, dG_s(\theta),$$

$$G_s = \textstyle\sum_k \pi_{sk}\, \delta_{\theta_k^* - \mu_s},$$

$$G_s^* \mid G^*, M_s \sim \text{DP}(M_s, G^*), \qquad G^* \mid \gamma \sim \text{DP}(\gamma, H), \quad H = \mathcal{N}(0, \sigma_\theta^2),$$

$$\gamma \sim \text{Gamma}(a_\gamma, b_\gamma), \qquad M_s \sim \text{Gamma}(a_M, b_M), \qquad \sigma_s^2 \sim \nu\lambda/\chi_\nu^2.$$

# References


- Chipman, H.A., George, E.I., McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1), 266–298.
- Hahn, P.R., Murray, J.S., Carvalho, C.M. (2020). Bayesian regression tree models for causal inference. *Bayesian Analysis*, 15(3), 965–1056.
- Robins, J.M. (2000). Sensitivity analysis for selection bias and unmeasured confounding. *IMA Volumes in Mathematics*, 116, 1–94.
- Neuenschwander, B., et al. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1), 5–18.
- Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M. (2006). Hierarchical Dirichlet processes. *JASA*, 101(476), 1566–1581.
- Brathovde, M., Putter, H., Valberg, M., Post, R.A.J. (2024). The causal interpretation of acceleration factors. *arXiv:2409.01983*.