Uncertainty Calibration and Exemplar Identification for Heterogeneous Treatment Effects with Individualized Bayesian Causal Forests (iBCF)

Jennifer Starling, Dan Thal, Lauren Vollmer, Irina Degtiar, Erin Lipman, Peter Mariani, and Mariel Finucane

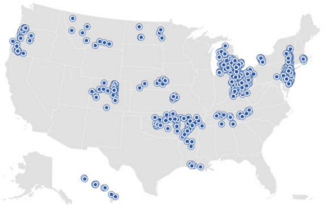
Mathematica

ISBA
July 1, 2022

Motivation: Improving primary care

The Center for Medicare and Medicaid Innovation (CMMI) is a group in the Center for Medicare and Medicaid Services (CMS), with a bipartisan charge to decrease primary care spending and improve care for patients.

- CMMI designs **alternative payment models** to reward healthcare providers for delivering high-quality, cost-efficient care
- **Practices** voluntarily participate in these plans
- **Primary Care First** (PCF) program is currently underway



Source: Centers for Medicare & Medicaid Services

Evaluating alternative payment models

Goals:

- Determine if the program worked overall (✓)
- Assess if the program worked for specified subgroups of interest (✓)
- Evaluate how well the program worked for each practice
    - With appropriate uncertainty
    - Well-calibrated data-driven subgroup analysis
    - Reward practices that perform better than expected

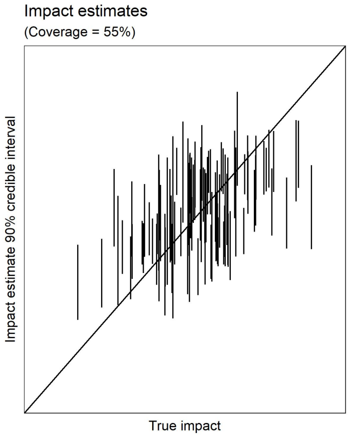
Practice impacts as potential outcomes

We can frame practice impacts in the potential outcomes framework. Let

- $y_i(1)$ be a practice's outcome under the program
- $y_i(0)$ be a practice's outcome absent participation

The practice-level impact is $y_i(1) - y_i(0)$.

Challenge 1: Uncertainty calibration for practice impact estimates



From Dorie (2019) on the 2016 ACIC data competition:

- ”Methods that flexibly model the response surface perform better overall than methods that fail to do so”
- ”Good coverage was difficult for most methods to achieve even when bias was low...we don’t feel like we have strong advice about how to optimize this aspect of performance”

Challenge 2: Identifying exemplar practices

A "high-performing" practice is one who

- Reduces expenditures
- Improves various outcome measures

Many unmeasured factors contribute to a practice's outcomes and response to payment programs.

- Energy and enthusiasm of practitioners
- Community support
- Participation in other programs
- Staff turnover

Data generating process

We estimate key data characteristics from real Medicare data, such as σ_y, ICC, and practice sizes.

We use these to simulate data that shares important features of the real data, such as

- Error variance
- Practice size

We generate non-linear control and impact functions using Gaussian covariates, scaled to mimic our real data. We sample bivariate practice-level random effects.

Introducing notation

Our data consists of n observations, indexed by i, where each observation represents a primary care practice.

Let

- y_i be the response (ex. expenditures, outcomes)
- z_i be a binary treatment indicator
- x_i are a vector of covariates for observation i
- $\pi(x_i)$ are propensity score estimates
- w_i are inverse practice sizes, which act as weights

BART model

BART is a Bayesian 'sum-of-trees' model introduced by Chipman, George, and McCulloch (2010). The BART model statement is:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$f(\mathbf{x}) = \sum_{j=1}^{m} g(\mathbf{x}, T_j, M_j = \{\mu_{j1} \ldots \mu_j l\})$$

We can also think of g as a basis function parameterized by the binary tree defined by (T_j, M_j).

BART prior is composed of priors on σ^2, terminal node values μ_{jl}, and tree structures T_j.

Bayesian Causal Forests (BCF)

Models response surface as the sum of two BART fits.

$$y_i = \underbrace{\mu(x_i, \pi(x_i)) + \tau(x_i) z_i}_{f(x_i)} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \frac{\sigma^2}{w_i}\right)$$

Using the potential outcomes framework, the treatment effect is

$$\tau(x_i) = f(x_i, 1) - f(x_i, 0)$$

Causal assumptions for Bayesian Causal Forests

- No interference between observations
- No unmeasured confounders
- Enough overlap to estimate impacts everywhere in covariate space

Under these conditions, $E\left[y_i(z) \mid t_i, x_i\right] = E\left[y_i \mid x_i, z_i = z\right]$, so we can express the causal estimand as:

$$\tau(t_i, x_i) = E\left[y_i \mid t_i, x_i, z_i = 1\right] - E\left[y_i \mid t_i, x_i, z_i = 0\right]$$

Benefits and shortcomings of Bayesian Causal Forests

Benefits:

- Deconfounding, through flexible modeling and less shrinkage of control covariates
- De-noising, through Bayesian shrinkage of impact estimates
- Flexible tree model tailored for learning impact heterogeneity
- Inclusion of propensity score estimates in control fit to mitigate bias

Shortcomings:

- Under-coverage
- Impact estimates for each practice are solely determined by x's

Introducing the iBCF model

$$y_i = \mu(x_i, \pi(x_i)) + \tau(x_i)z_i + \mathbf{u_i(1 - z_i) + (u_i + v_i)z_i} + \epsilon_i$$

$$\epsilon_i \sim N\left(0, \frac{\sigma^2}{w_i}\right)$$

Let

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}\right)$$

where

- u_i are random effects unrelated to treatment
- v_i are random impacts

The impact for practice i is $\tau(x_i) + v_i$.

iBCF is designed for weighted data

- For simplicity, write $\mathrm{Var}(u_i + v_i) = \sigma_u + \sigma_v + 2\rho\sigma_u\sigma_v = \sigma_b^2$
- The likelihood factorizes

$$y \sim \prod_{i|z_i=0}^{n_C} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_u^2\right) \times \prod_{i|z_i=1}^{n_T} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_b^2\right)$$

Weights allow us to separately identify σ_y versus the variance from the random effects.

- σ_y is the portion of the error variance that goes to zero as practice size $\rightarrow \infty$
- Even absent uncertainty, there is still remaining heterogeneity among the practices

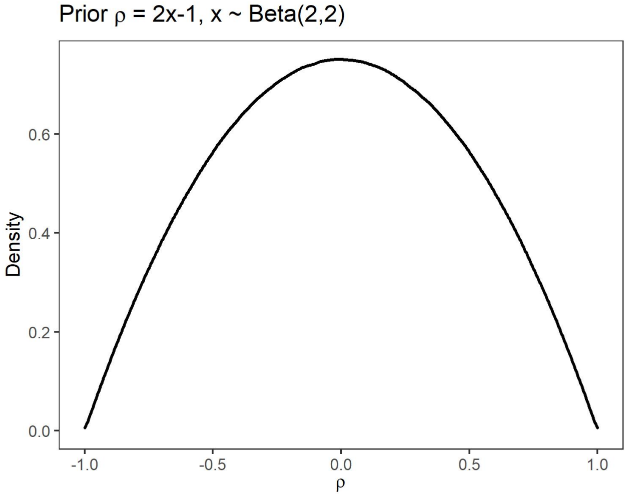
iBCF priors

We choose priors as follows:

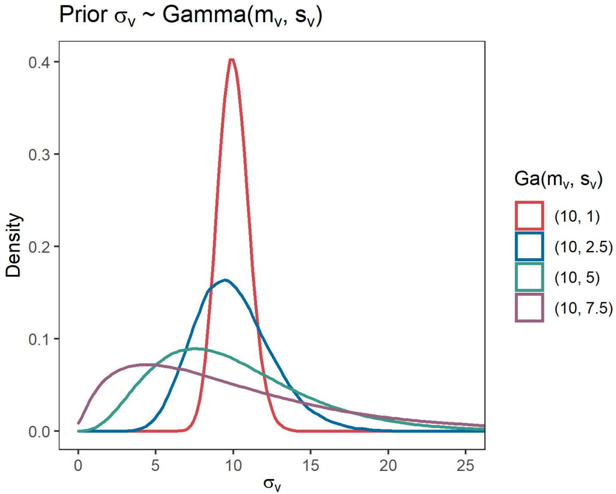
$$\begin{aligned}
\mathsf{Var}(u_i) = \sigma_u &\quad \sim C^+(1) \\
\mathsf{Var}(u_i + v_i) = \underbrace{\sigma_u + \sigma_v + 2\rho\sigma_u\sigma_v}_{\sigma_b} &\quad \sim C^+(1) \\
x = \frac{\rho - 1}{2} &\quad \sim Beta(2, 2) \\
\sigma_v &\quad \sim \mathsf{Gamma}(m_v, s_v)
\end{aligned}$$

We elicit the Gamma prior mean and sd for σ_v, which is not identified by the data.

Prior for $\rho = \text{Corr}(u_i, v_i)$



Prior for $\sigma_v = \text{Var}(v_i)$



Fitting the iBCF model

We fit iBCF using a modified version of BCF's backfitting algorithm.

Updating variance components:

1. MH step to draw σ_y; no longer conjugate
2. Posterior draws of $\sigma_u^2 = \mathsf{Var}(u_i) \mid y_C$ and $\sigma_b^2 = \mathsf{Var}(u_i + v_i) \mid y_T$
3. Posterior draws of ρ and σ_v (posterior=prior)

Drawing practice random effects:

1. Posterior draw of pair $(u_i, u_i + v_i)$
2. Calculate posterior draw for v_i as $v_i = (u_i + v_i) - u_i$

Data-driven hyperparameter tuning for $\sigma_v \sim Ga(m_v, s_v)$

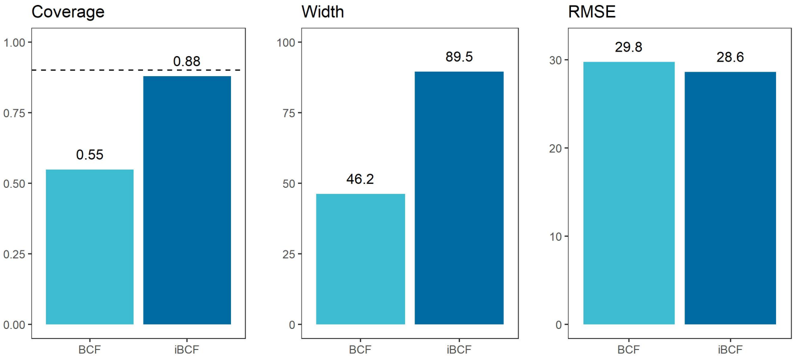
Let $\text{Var}(u_i + v_i) = \sigma_b^2$ and recall that the likelihood factorizes:

$$y \sim \prod_{i|z_i=0}^{n_C} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_u^2\right) \times \prod_{i|z_i=1}^{n_T} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_b^2\right)$$

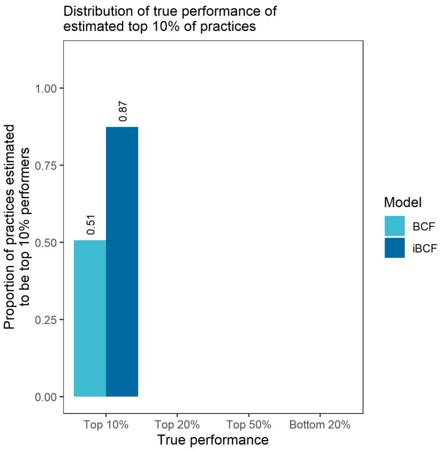
- Estimate residuals by fitting BART models to treated and control data
- Estimate σ_u and σ_b as intercepts from regressing $r_i^2 \sim (1/w_i)$ for treated and control observations (with positivity constraints)
- Solve the quadratic equation $\sigma_v^2 = \sigma_b^2 - \sigma_u^2 - 2\rho\sigma_u\sigma_v$ using $\hat{\sigma}_u$ and $\hat{\sigma}_b$ and a range of specified ρ values

We let $m_v = \hat{\sigma}_v$ and $s_v = 0.25 m_v$.

Practice-specific impact estimates: coverage, interval width, and RMSE



Identifying practices with exemplar outcomes

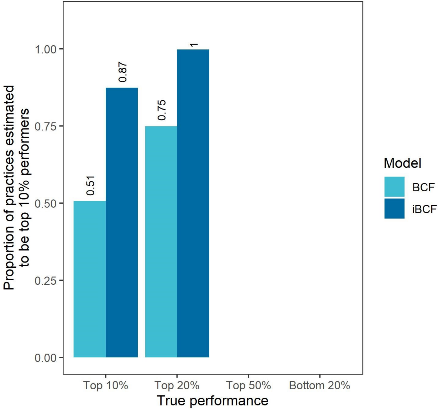


Of the practices we estimate to have outcomes in the top 10%,

- Only 51% have true top-10% outcomes under BCF
- 87% have true top-10% outcomes under iBCF

Identifying practices with exemplar outcomes

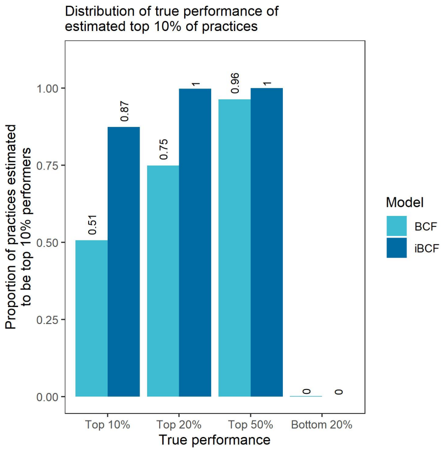
Distribution of true performance of estimated top 10% of practices



Of the practices we estimate to have outcomes in the top 10%,

- Only 75% are truly in the top 20% under BCF
- 100% are truly in the top 20% under iBCF

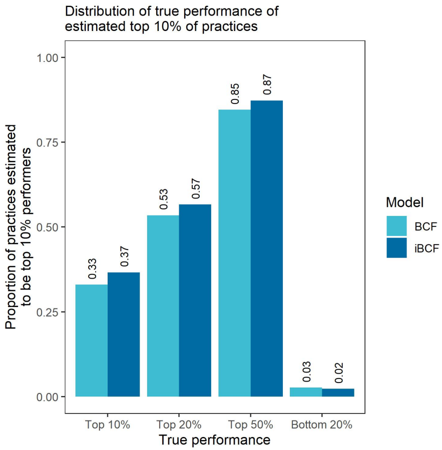
Identifying practices with exemplar outcomes



Of the practices we estimate to have outcomes in the top 10%,

- No practices are in the bottom 20% for either method.
- BCF does have 4% of practices in the bottom half

Identifying practices with exemplar impacts



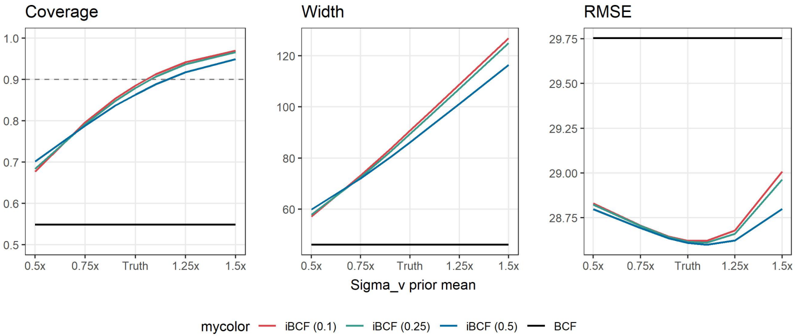
Of the practices we estimate to have impacts in the top 10%,

- 37% are truly in the top 10% under iBCF
- 33% are truly in the top 10% under BCF

Modest gains due to fairly small proportion of total impact variability accounted for by v_i in Medicare setting.

Sensitivity analysis for $\sigma_v \sim Ga(m_v, s_v)$ hyperparameters

- As $\sigma_v \to 0$, iBCF reverts towards BCF specification.
- BCF implies the prior $\sigma_v = 0$.



Summary

Novel approach for estimating observation-level impacts

- Improved uncertainty calibration
- Successfully identify top-performing practices
- iBCF performed well in the recent ACIC data challenge

Results are sensitive to the choice of hyperpriors.

- Refining our data-driven method assists with tuning
- Provides a more sensible prior than assuming $\sigma_v = 0$

References

- Hahn, P.R., Murray, J. S. and Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. Bayesian Analysis, 15 (3), 965–1056.
- Chipman, H.A. and George, E.I. and McCulloch, R.E. (2010). BART: Bayesian Additive Regression Trees. Annals of Applied Statistics, 4(1) 266-298.
- Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. Journal of Computational and Graphical Statistics, 20 (1), 217–240.

Appendix

Data generating process - details

We estimate the following quantities from real Medicare data.

- Quantiles for practice sizes
- σ_y^2, residual variance
- ICC (within-practice variance / total variance)
- σ_u estimated from ICC $= \frac{\sigma_u^2}{\sigma_u^2+\sigma_y^2}$
- σ_v estimated using our pre-regression technique
- $\text{Var}(\tau(x_i))$ fitting BCF to real data

Data generating process - details

We then generate data using our estimated values as follows:

- Weights drawn from real practice size quantiles
- Covariates drawn from standard normal distribution
- $\mu(x_i, \pi_i)$ non-linear function of covariates, with some (measured) confounders
- $\tau(x_i)$ Non-linear function of covariates, scaled to mimic Var($\tau(x_i)$)
- Draw random effects u_i, v_i from MVN with specified ρ

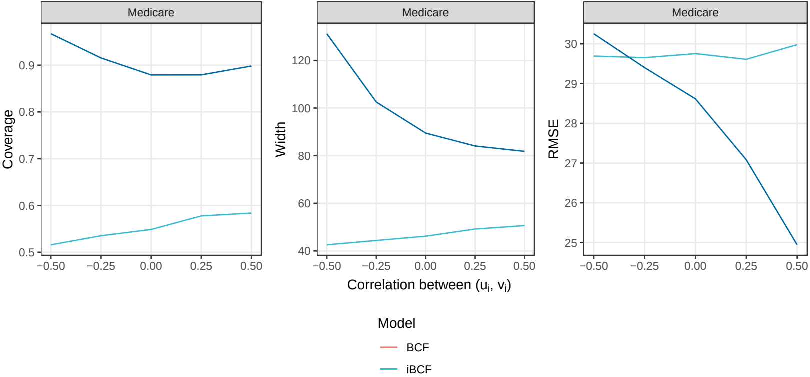
Causal assumptions for Bayesian Causal Forests - Details

- SUTVA (Stable Unit Treatment Value Assumption)
  - No interference between units, i.e.
  - The response of an observation depends only on its treatment, not on the treatment of other observations around it
- Strong ignorability, consisting of two conditions:
  - No unmeasured confounders: $(y_i(0), y_i(1)) \perp (z_i \mid t_i, X_i)$
  - Enough overlap to estimate treatment effects everywhere in covariate space: $0 < Pr(z_i = 1 \mid t_i, x_i) < 1$

Under these conditions, $E\left[y_i(z) \mid t_i, x_i\right] = E\left[y_i \mid x_i, z_i = z\right]$, so we can express the causal estimand as:

$$\tau(t_i, x_i) = E\left[y_i \mid t_i, x_i, z_i = 1\right] - E\left[y_i \mid t_i, x_i, z_i = 0\right]$$

Sensitivity to ρ in the simulated data



Data-driven hyperparameter tuning for $\sigma_v \sim Ga(m_v, s_v)$

Let $\text{Var}(u_i + v_i) = \sigma_b^2$ and recall that the likelihood factorizes:

$$y \sim \prod_{i|z_i=0}^{n_C} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_u^2\right) \times \prod_{i|z_i=1}^{n_T} N\left(f(x_i), \frac{\sigma_y^2}{w_i} + \sigma_b^2\right)$$

1. Estimate σ_u using control observations ($\sigma_i^2 = \frac{\sigma_y^2}{w_i} + \sigma_u^2$).
   - Fit a BART model to the control observations; let r_i^2 be the estimated squared residuals.
   - Fit linear model (I(r_i^2) $\sim$ 1/size)). The intercept gives $\hat{\sigma}_u^2$. (Use glmnet with λ=0 and lower.limit=0 for positivity constraints)
2. Repeat for treated observations to estimate σ_b ($\sigma_i^2 = \frac{\sigma_y^2}{w_i} + \sigma_b^2$).
3. Solve quadratic equation to estimate σ_v^2 for a range of ρ values.
   - $\sigma_v^2 = \sigma_b^2 - \sigma_u^2 - 2\rho\sigma_u\sigma_v$
   - Let $m_v = \hat{\sigma}_v^2$ and $s_v = 0.25m_v$.

Data-driven hyperparameter tuning for $\sigma_v \sim Ga(m_v, s_v)$

The quadratic equation is derived as follows. Let $b_i = u_i + v_i$.

$$\begin{aligned}
\sigma_v^2 &= \mathsf{Var}(v_i) \\
&= \mathsf{Var}(b_i - u_i) \\
&= \sigma_u^2 + \sigma_b^2 - 2\mathsf{Cov}(u_i, b_i) \\
&= \sigma_u^2 + \sigma_b^2 - 2\mathsf{Cov}(u_i, u_i + v_i) \\
&= \sigma_u^2 + \sigma_b^2 - 2(\sigma_u^2 + \mathsf{Cov}(u_i, v_i)) \\
&= \sigma_u^2 + \sigma_b^2 - 2(\sigma_u^2 + \rho\sigma_u\sigma_v) \\
&= \sigma_u^2 + \sigma_b^2 - 2(\sigma_u^2 + \rho\sigma_u\sigma_v) \\
&= \sigma_b^2 - \sigma_u^2 - 2\rho\sigma_u\sigma_v
\end{aligned}$$

Fitting the BART model using Bayesian Backfitting

The BART model is fit using an iterative MCMC called 'Bayesian Backfitting.'

- Trees T_j for $j \in \{1, \ldots, m\}$ are updated one by one, using residuals from fits of other trees
  - Each tree is drawn using a MH step to select a node and propose a birth or death move
- Tree leaves μ are drawn from conditional posterior
- Error variance σ is drawn from conditional posterior

The μ_{jl} and σ updates are easy; priors are conjugate.