



A FLEXIBLE APPROACH TO PREDICTIVE BIOMARKER DISCOVERY

Philippe Boileau¹, Nina Ting Qi², Mark J. van der Laan¹, Sandrine Dudoit¹, Ning Leng²

¹University of California, Berkeley; ²Genentech Inc.

Background

This is the background.

Variable Importance Parameter

Consider n identically and independently distributed (i.i.d) full-data random vectors $X = (W, A, Y^{(0)}, Y^{(1)}) \sim P_X$.

- W : A p -length random vector of centered pretreatment biomarkers with nonzero variance.
- A : A random binary indicator of treatment assignment.
- $Y^{(0)}, Y^{(1)}$: Continuous potential outcomes under assignment to the control and treatment allocations, respectively.

Our causal variable importance parameter is $\Psi^F(P_X) = (\Psi_1^F(P_X), \dots, \Psi_p^F(P_X))$, where

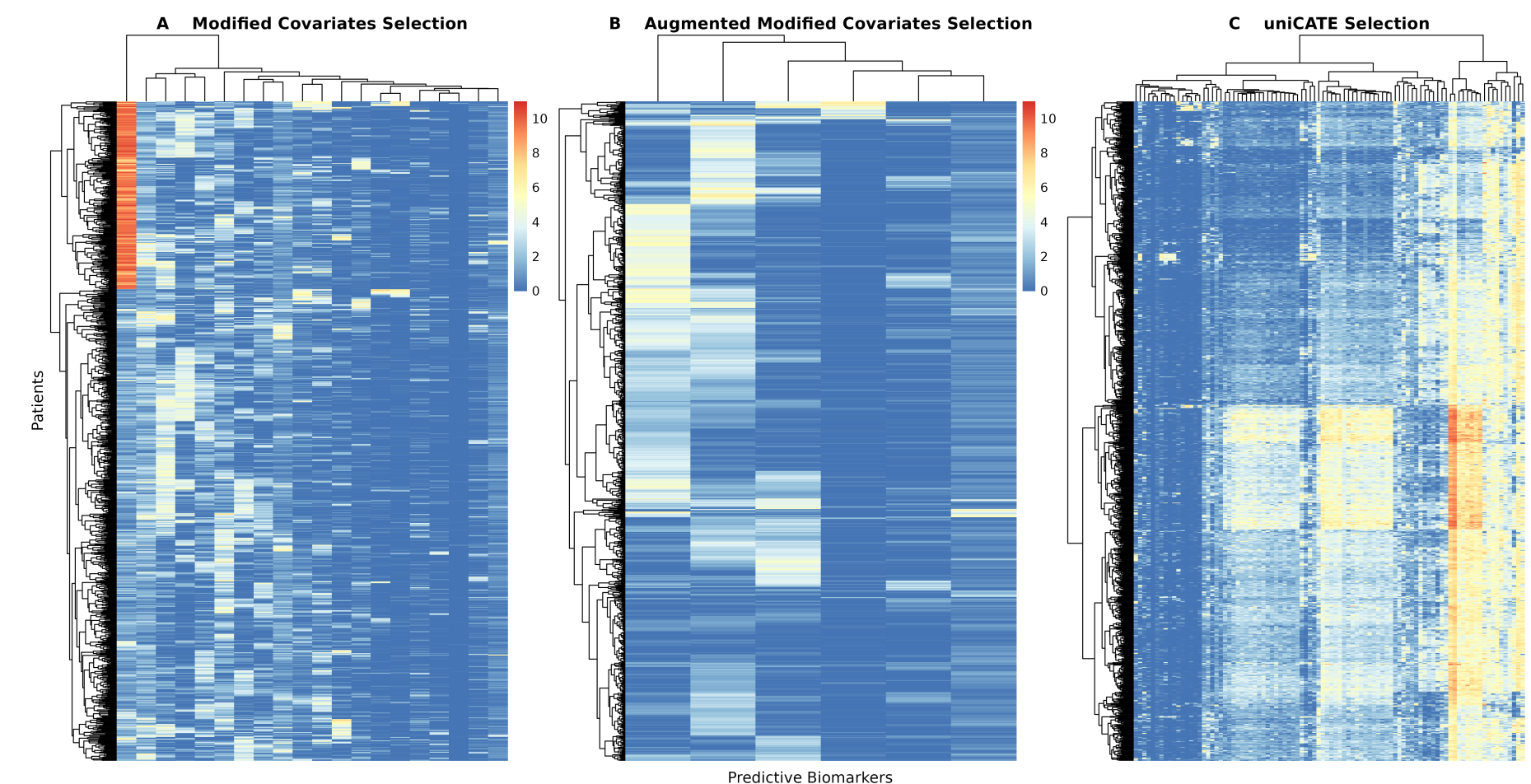
$$\Psi_j^F(P_X) \equiv \frac{\mathbb{E}_{P_X} \left[\left(Y^{(1)} - Y^{(0)} \right) W_j \right]}{\mathbb{E}_{P_X} \left[W_j^2 \right]}.$$

Given access instead to n i.i.d. censored random observations $O = (W, A, Y)$ where $Y = AY^{(1)} + (1 - A)Y^{(0)}$, $\Psi^F(P_X)$ is identified under the assumptions of unmeasured confounding and positivity by $\Psi(P_0) = (\Psi_1(P_0), \dots, \Psi_p(P_0))$. Here,

$$\Psi_j(P_0) \equiv \frac{\mathbb{E}_{P_0} \left[(\bar{Q}_0[A = 1, W] - \bar{Q}_0[A = 0, W]) W_j \right]}{\mathbb{E}_{P_0} \left[W_j^2 \right]},$$

where $\bar{Q}_0[A, W] = \mathbb{E}_{P_0}[Y|A, W]$.

Application to IMmotion 150 and 151



Inference

Let $g_0(W) = \mathbb{P}[A = 1|W]$, and let \hat{g} and \hat{Q} be estimators of g_0 and \bar{Q}_0 , respectively. Define the Augmented Inverse Probability Weighted outcome difference as

$$T(O) \equiv \left(\frac{I(A = 1)}{\hat{g}(W)} - \frac{I(A = 0)}{1 - \hat{g}(W)} \right) (Y - \hat{Q}(A, W)) + \hat{Q}(1, W) - \hat{Q}(0, W).$$

We derive from the efficient influence function of $\Psi_j(P_0)$, $D_j(O)$, the double-robust one-step estimator

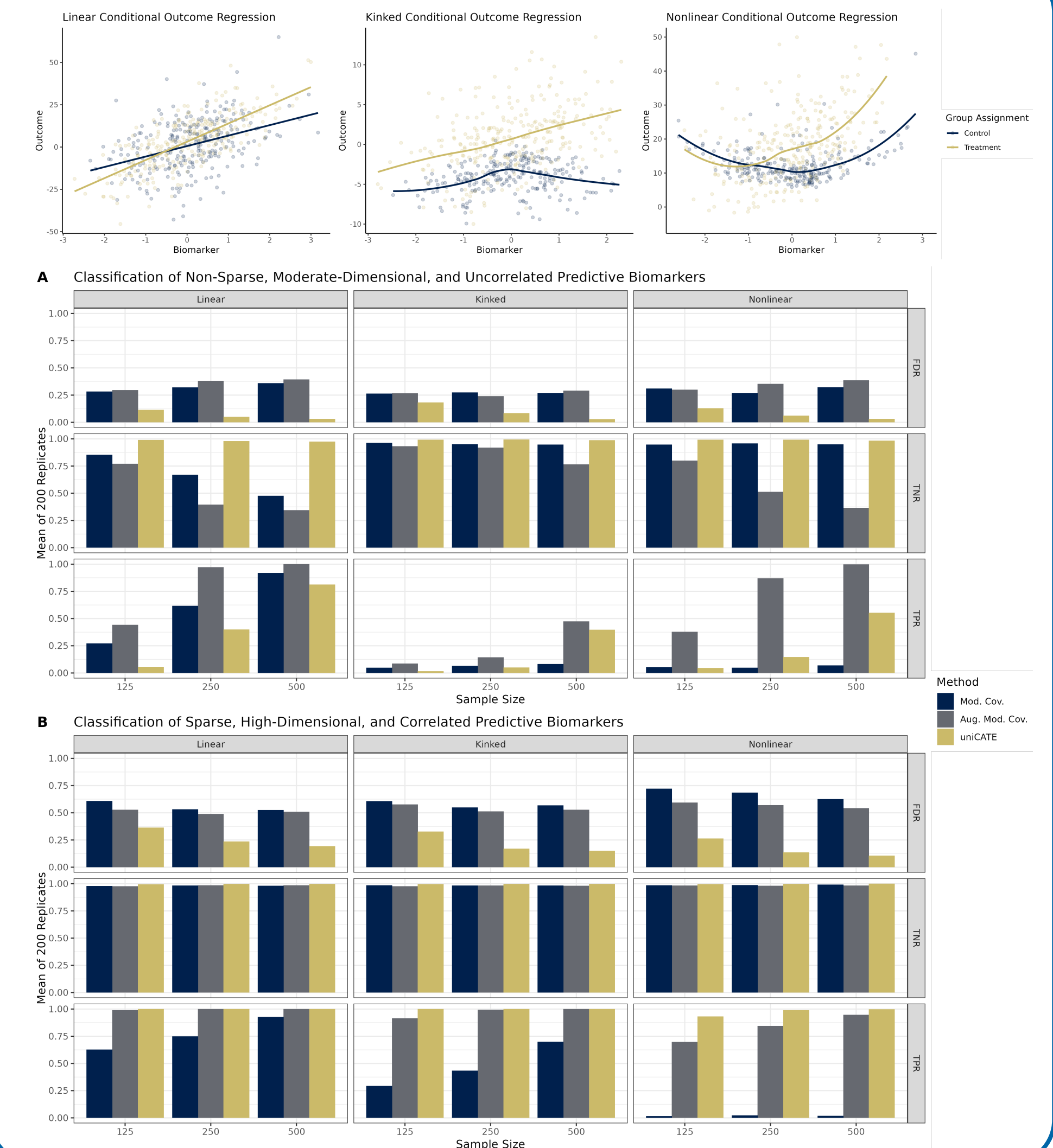
$$\hat{\Psi}_j(P_n) \equiv \frac{\sum_{i=1}^n T(O_i) W_{ij}}{\sum_{i=1}^n W_{ij}^2},$$

where $\sum_i W_{ij} = 0$ for all j and P_n is the empirical distribution.

If \hat{g} and \hat{Q} are trained using sample splitting techniques, and we assume that $\|\hat{g} - g_0\|_2 \|\hat{Q} - \bar{Q}_0\|_2 = o_p(n^{-1/2})$, then

$$\sqrt{n} \left(\hat{\Psi}_j(P_n) - \Psi_j(P_0) \right) \xrightarrow{D} N(0, \mathbb{V}_{P_0}[D_j(O)]).$$

Randomized Control Trial Simulations



Conclusions

Predictive biomarker discovery benefits from formal statistical inference procedures that control Type-I error rates. This estimator is implemented in the **uniCATE** R package available at github.com/insightsengineering/uniCATE.

Funding

PB gratefully acknowledges the support of the National Science and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies.

References

PB gratefully acknowledges the support of the National Science and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies.