1. Estimating the risk difference as estimand of interest. Rather than looking at the ratio of potential outcomes, it is often the case we want to investigate the difference in the expected value of each, i.e. we can look at *risk differences*:

$$\text{Risk Difference} \rightarrow \delta \equiv \mathrm{E}(B^1) - E(B^0)$$

In our framework, following similar reasoning as in section *Modular sensitivity analysis with machine learning*in the main file, risk differences can be defined as

$$\Delta(\mathbf{x}_i) = \int_{\mathbb{R}} \Phi\big(b_1(\mathbf{x}) + u\big) f(u)\mathrm{d}u - \int_{\mathbb{R}} \Phi\big(b_0(\mathbf{x}) + u\big) f(u)\mathrm{d}u$$

The sample average risk difference (ARD) is therefore $\frac{1}{n}\sum_{i=1}^{n}\Delta(\mathbf{x}_i)$. In the case of the audit data, the average risk difference refers to percentage point difference in going bankrupt after receiving a going concern. We esimate the risk difference using our methodology as well as the bivariate probit with endogenous regressor model, (described in equation(4) in the main file), to the audit data. Specifically, we used the same covariates as we used when fitting monotone bart, used the bankruptcy indicator as the binary outcome, and whether or not a going concern was issued as the "treatment" indicator. Using our methodology, results for estimating risk differences on the audit data are presented in table 2.

γ	ρ	ARD true	ARD est	IRD cor	IRD RMSE
1.00	0.25	0.29	0.29	0.89	0.05
1.75	0.25	0.47	0.47	0.96	0.04
2.50	0.25	0.58	0.57	0.97	0.05
1.00	0.40	0.29	0.29	0.90	0.05
1.75	0.40	0.47	0.46	0.94	0.06
2.50	0.40	0.58	0.55	0.96	0.09
1.00	0.60	0.29	0.29	0.90	0.05
1.75	0.60	0.47	0.46	0.95	0.05
2.50	0.60	0.58	0.57	0.98	0.04
1.00	0.80	0.29	0.27	0.91	0.05
1.75	0.80	0.47	0.44	0.95	0.05
2.50	0.80	0.58	0.55	0.98	0.05

TABLE 1

We simulate from the bivariate probit with 25,000 observations and deploy our methodology. cor refers to the correlation between predicted and true for the average risk difference (ARD), and the rmse is the root mean square error.

When fitting the bivariate probit regression, parameters are estimated using maximum likelihood. Our ARD from this regression was 2.90 percentage points and a mean risk ratio of 2.82 with 95% CI (1.33, 6.04). Note, we used a regression spline approach to smooth our numeric covariates, where the smooth term for each covariate is made of basis functions, see Marra and Radice [2011] for details. This gives us additional flexibility in fitting the model.

It was stressed in the main document how using BART with a monotonicity constraint improves our estimation of the ICRR, but the improvement is even more pronounced when studying risk differences. In figure 1, we look at the comparison of IRD (individual risk difference) estimates from data generated by the bivariate probit, with the left hand side of our system of equation probabilites estimated with BART and monotone BART. This is a similar plot to figure 14 in the main file, but with a different estimand of interest.

Distribution of $f(u)$	ARD (%)	ARD post (%)	mean B_1 (%)	95% Credible interval for ARD (%)
N$(0, \sigma = 0.1)$	8.89	8.91	10.3	$(6.95, 11.2)$
N$(0, \sigma = 0.5)$	3.62	3.76	5.29	$(2.78, 4.97)$
N$(0, \sigma = 1)$	0.41	0.63	2.57	$(0.37, 0.97)$
Shark $q = 0.25$, $s = 0.5$; $\sigma = 1.05$	0.11	0.24	2.44	$(0.13, 0.40)$
Shark $q = 0.75$, $s = 1.25$; $\sigma = 0.88$	2.39	2.57	4.30	$(1.82, 3.52)$
"Right Bump" $\sigma = 0.48$	1.98	2.17	4.10	$(1.78, 2.69)$
98% peak $\sigma = 0.29$	6.85	7.08	8.62	$(5.41, 9.06)$
90% peak $\sigma = 0.64$	1.85	2.03	4.02	$(1.68, 2.51)$

TABLE 2

The reduced form probabilities (equation (15) in the main file) were estimated using BART with a monotonicity constraint on the going concern variable. We further require $b_1(\mathbf{x}) > b_0(\mathbf{x})$ in the projection step. Posterior summaries based on 500 Monte Carlo samples. σ refers to the implied standard deviations of the different distributions.

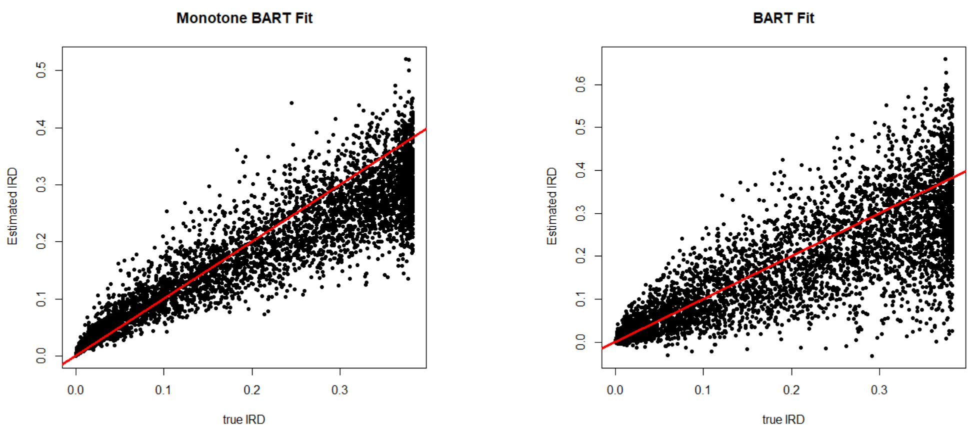


Fig 1: Plots of expected individual risk differences (IRD) vs our estimates, i.e. a plot comparing the difference in potential outcomes from the bivariate probit model $(\Phi(\alpha_0 + \alpha_1 \mathbf{x}_i + \gamma) - \Phi(\alpha_0 + \alpha_1 \mathbf{x}_i))$ versus our estimate. In the DGP, $\rho = 0.25, \gamma = 1$. The monotone BART correlation between τ and $\hat{\tau}$ is 0.928 and for BART is 0.826

2. Bivariate probit simulation study. Table 3 shows the results when fitting the bivariate probit regression with a maximum likelihood estimate to the simulated bivariate probit data. Unsurprisingly, this performs well, with the caveat that we require large N (N $= 100,000$) to get these impressive results. We simulated the samples from the bivariate probit model of the main document, where we sum 5 uniform(-1,1) $\mathbf{x}_i$ covariates each with the same β_1 and α_1 coefficients respectively. We set $\beta_0 = 0, \beta_1 = -0.2, \alpha_0 = -0.5, \alpha_1 = 0.7$ to generate reasonable number of going concerns and bankruptcies.

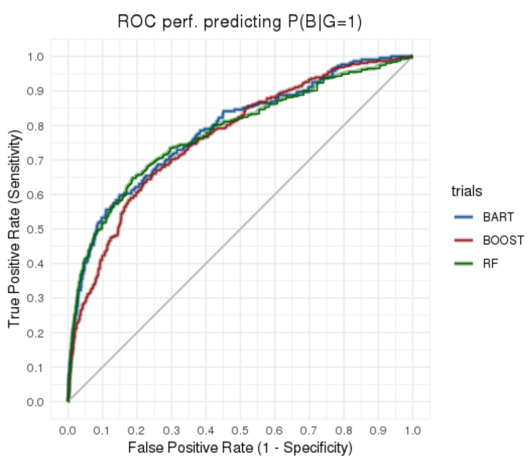
Table 1 shows the results when we simulate from the bivariate probit and fit with our methodology, with $f(u)$ assigned appropriately, only this time we are interested in the treatment effect. Our method does well here, with only N $= 25,000$ and $p = 5$.

ARD true	ARD est	IRD cor	IRD RMSE	ACRR true	ACRR est	ICRR cor	ICRR rmse	γ true	γ est.	ρ	ρ est.
0.23	0.24	0.97	0.02	2.24	2.07	1.00	0.24	1.00	0.77	0.25	0.37
0.46	0.46	0.99	0.02	4.32	3.89	1.00	0.87	1.75	1.62	0.25	0.31
0.58	0.57	1.00	0.02	6.24	5.40	0.99	2.34	2.50	2.38	0.25	0.30
0.26	0.26	0.99	0.01	2.42	2.27	1.00	0.22	1.00	0.85	0.40	0.47
0.46	0.46	0.99	0.02	4.33	3.89	1.00	0.90	1.75	1.63	0.40	0.45
0.57	0.56	0.99	0.02	6.14	5.13	1.00	2.83	2.50	2.34	0.40	0.46
0.28	0.28	0.99	0.01	2.57	2.46	1.00	0.17	1.00	0.92	0.60	0.63
0.47	0.47	1.00	0.01	4.51	4.24	1.00	0.63	1.75	1.70	0.60	0.61
0.59	0.58	1.00	0.01	6.41	5.89	0.99	1.67	2.50	2.45	0.60	0.61
0.31	0.31	1.00	0.00	2.79	2.80	1.00	0.02	1.00	1.02	0.80	0.80
0.47	0.47	1.00	0.01	4.51	4.26	1.00	0.56	1.75	1.70	0.80	0.81
0.58	0.58	1.00	0.01	6.31	5.56	0.99	2.25	2.50	2.41	0.80	0.82

TABLE 3

N=100,000. Fit the simulated bivariate probit with the bivariate probit regression. Validates the MLE of the bivariate probit regression performs well, as well as the validity of our data generation process, however required a large N to get accurate results. ARD refers to average risk difference, IRD individual risk difference. ACRR refers to average causal risk ratio, whereas ICRR is individual causal risk ratio.

3. Comparing machine learning methods for the observational data. Here we present our results from fitting the left hand side of equation(15) in the main file. In this section, we compare the performance in predicting the left hand side of equation(15) using various non-parametric “machine learning” tools. In particular, we compare using monotone BART (described earlier), random forests, and XGBoost. Referencing figure 2 actually tells two different tales. According to the table, all the methods perform relatively similarly by the auc metric, but the roc curve shows the methods provide different probability estimates.



	XGBoost	RanFor	BART
Case	AUC	AUC	AUC
B_1, G_0	0.76	0.77	0.78
B_1, G_1	0.79	0.79	0.79
G_1	0.88	0.89	0.89

Fig 2: Left: Area under curve of ROC plot, balanced 5-fold CV. Plot of ROC performance for predicting $\Pr\left(B \mid G=1, \mathbf{x}\right)$. Corresponds to shaded row in table on right. Right: case 1 is predicting bankruptcy when no going concern is issued, case 2 is when no concern issued, and case 3 predicts if concern is issued. All the methods are *similar*, with monotone BART seeming to be the top performer.

REFERENCES

G. Marra and R Radice. Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *Canadian Journal of Statistics*, 2011. 1