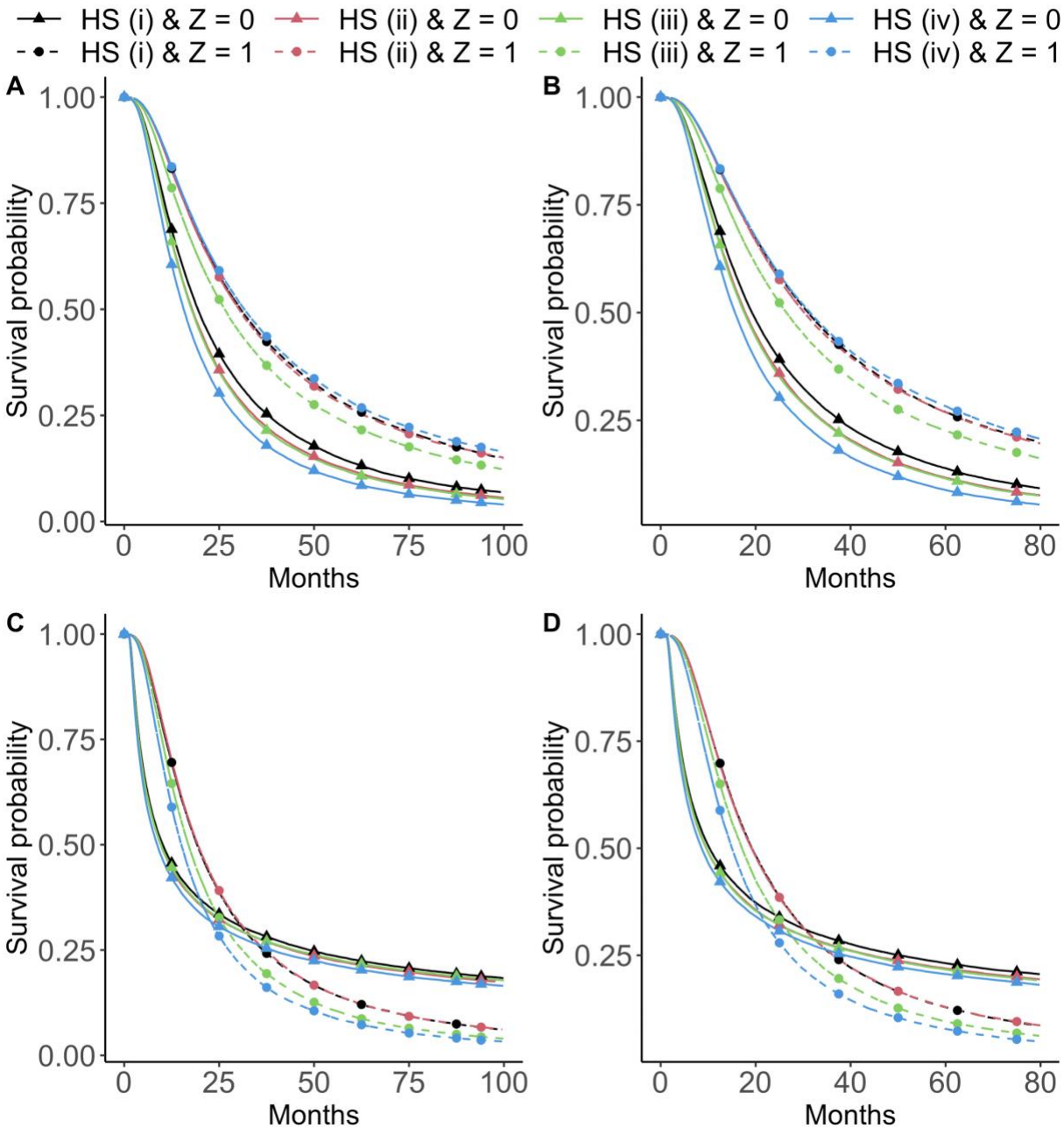
Supplemental Materials for "Estimating heterogeneous survival treatment effect in observational data using machine learning"

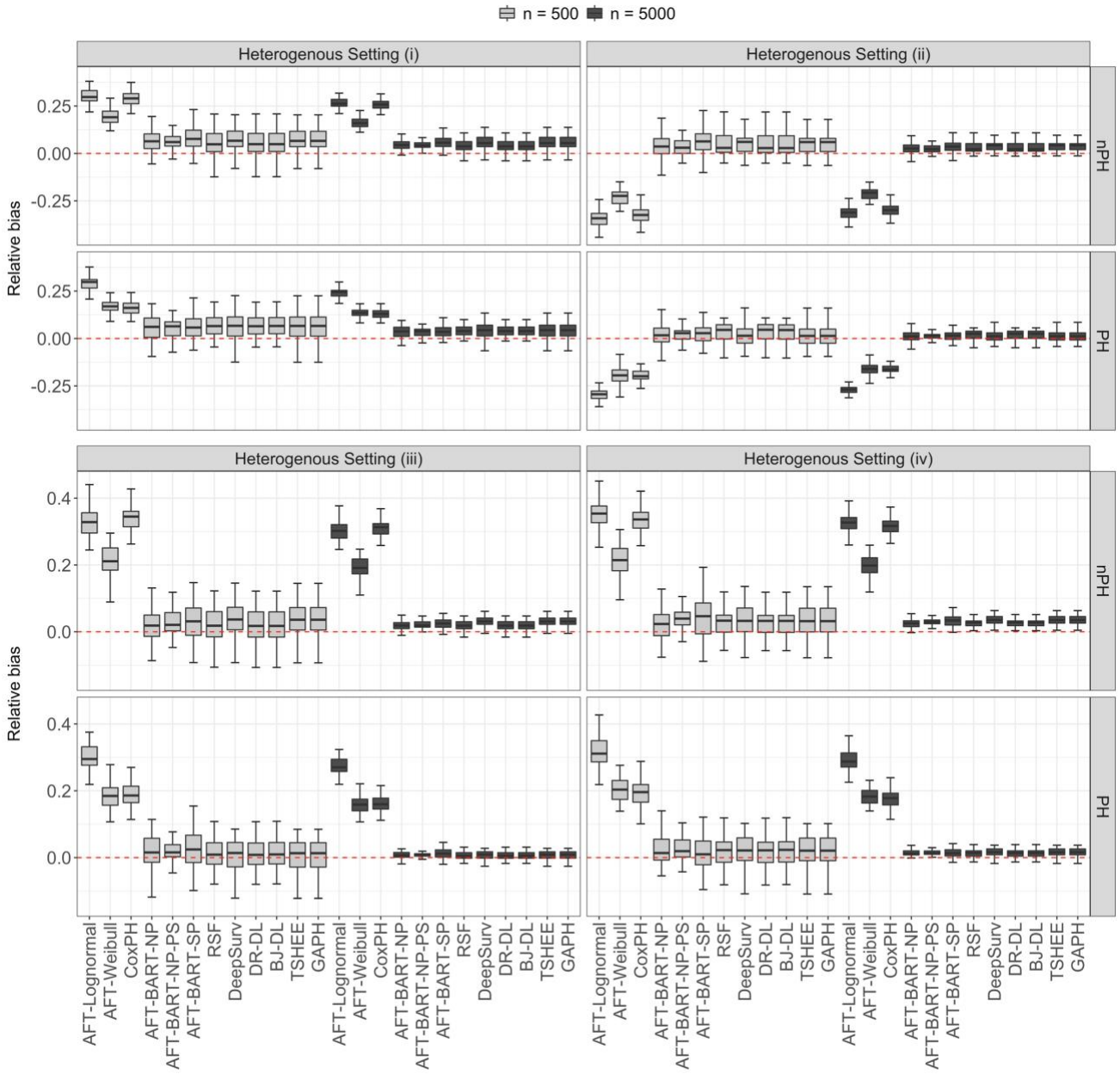
Liangyuan Hu, Jiayi Ji, Fan Li

Additional Figures and Tables

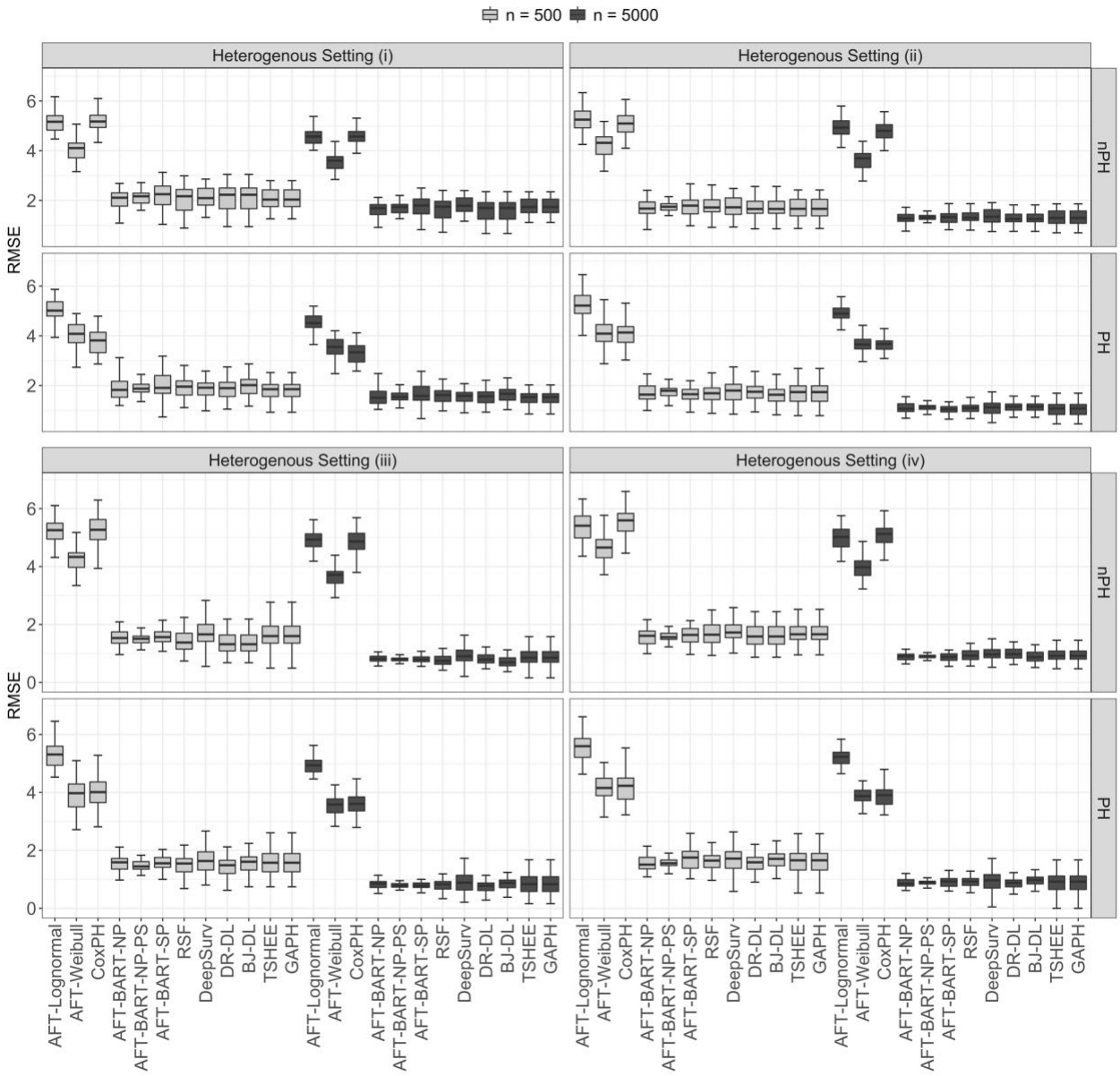


Web Figure 1: Kaplan-Meier survival curves for data simulated under our data generating processes. Each panel shows 8 survival curves for each treatment group and each of 4

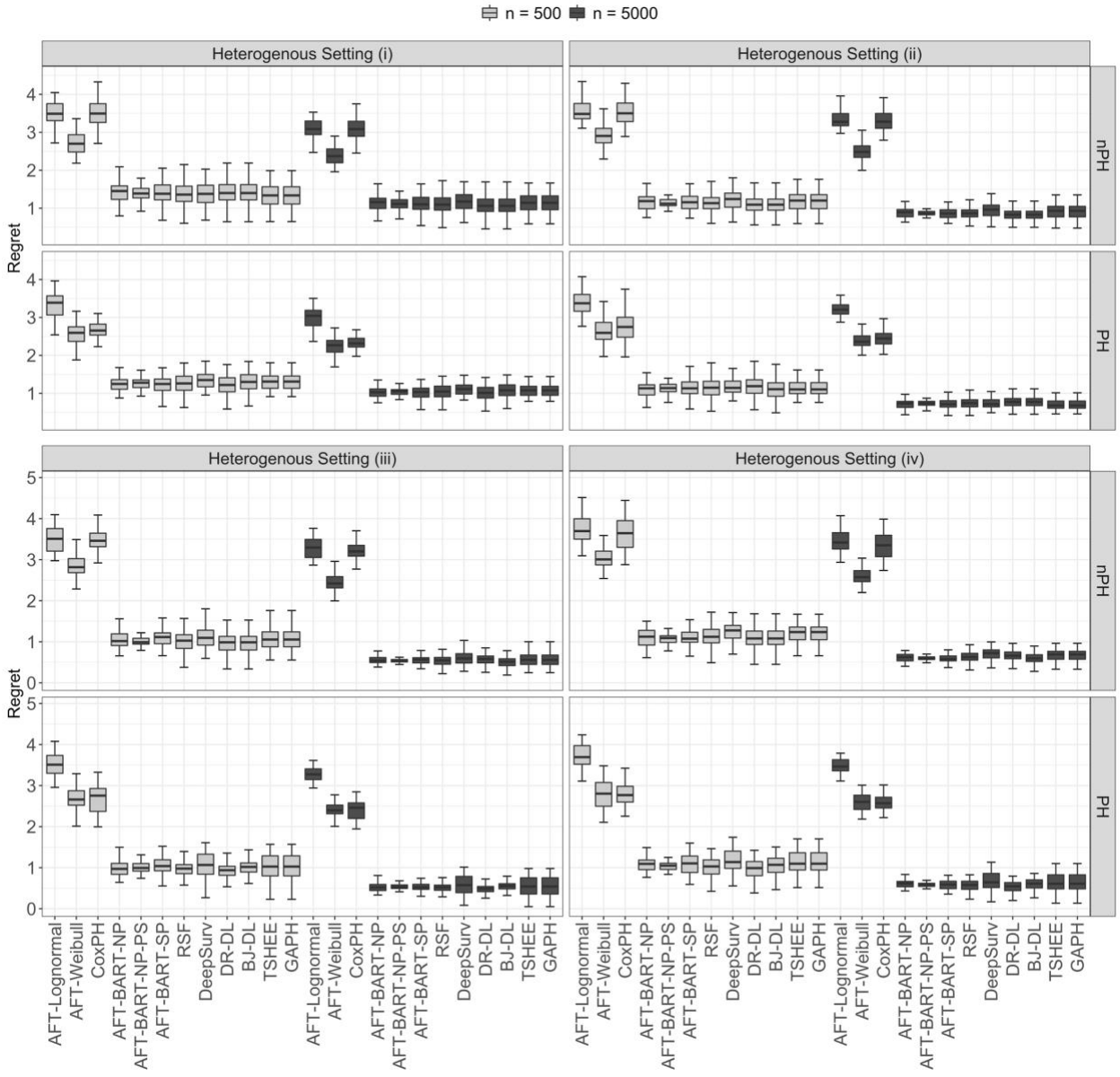
heterogenous settings. Panels A-D respectively represent scenarios corresponding to PH + 20% censoring, PH + 60% censoring, nPH +20% censoring and nPH +60% censoring.



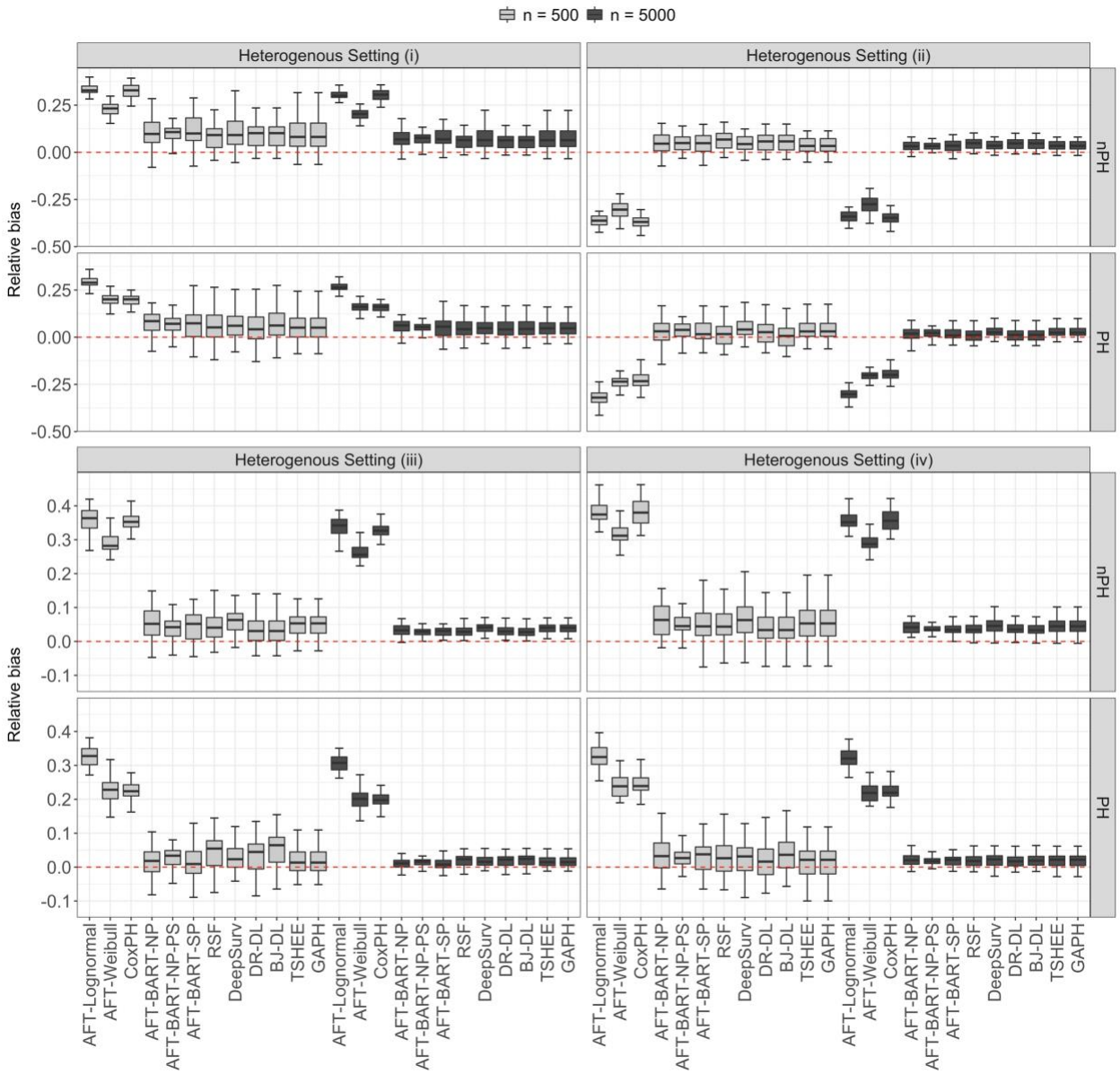
Web Figure 2: Relative biases among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 20% censoring. The relative bias for subgroup $k = 1, \ldots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



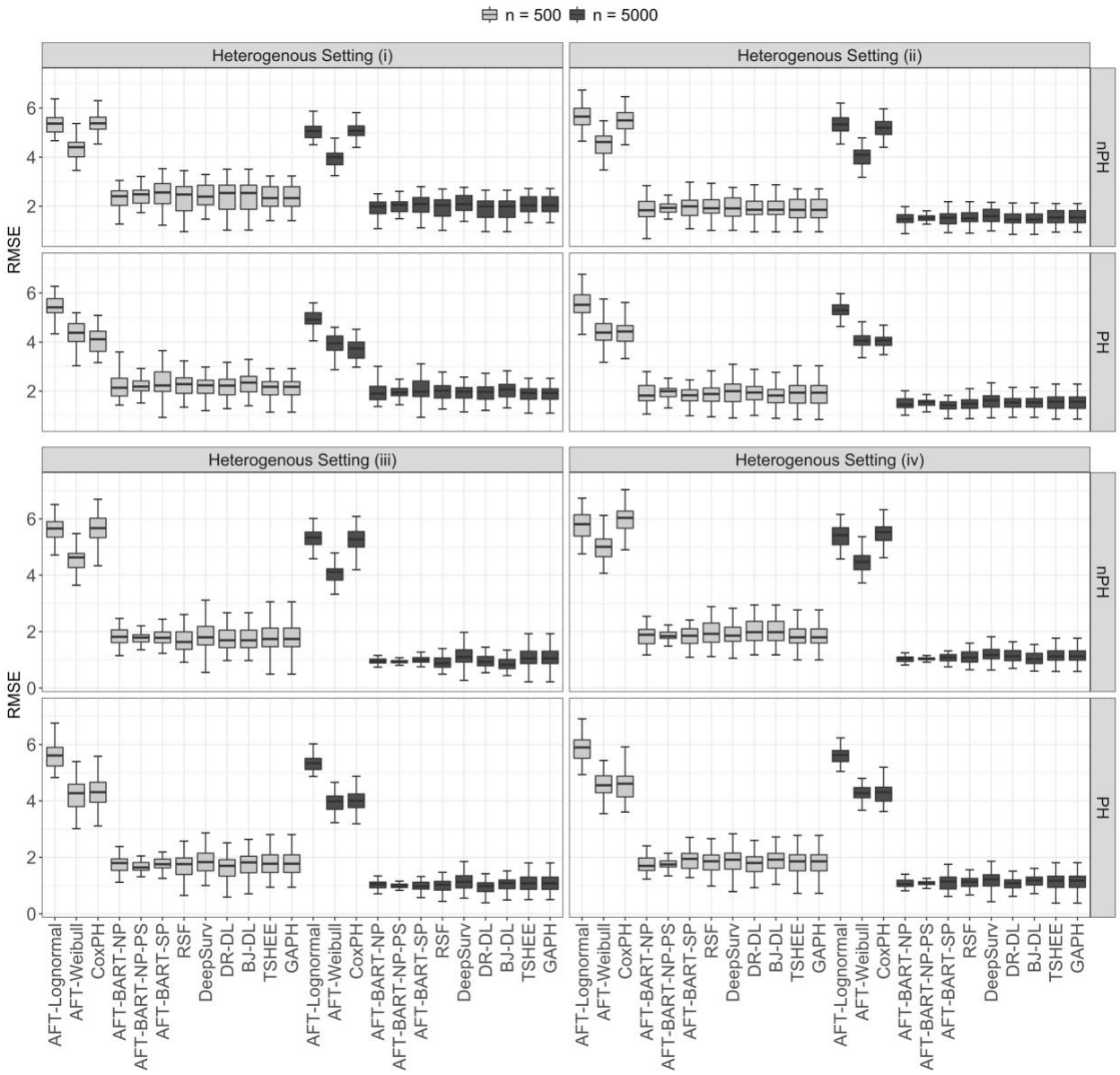
Web Figure 3: RMSE results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 20% censoring. The RMSE for subgroup $k = 1, \ldots, 50$ is defined by the root mean squared error between the estimated and true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



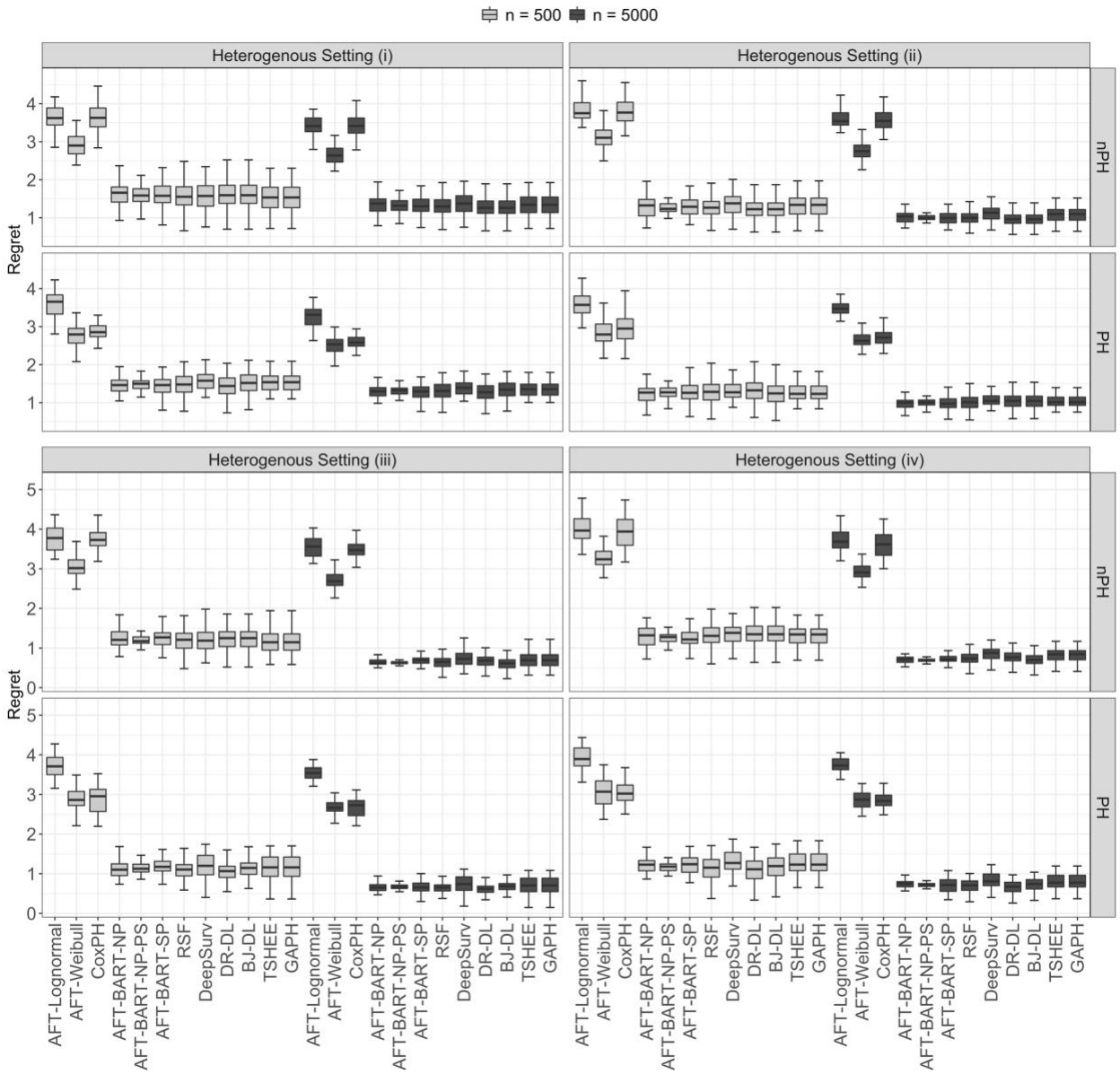
Web Figure 4: The regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 20% censoring. The RMSE for subgroup $k = 1, \ldots, 50$ is defined by the root mean squared error between the estimated and true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



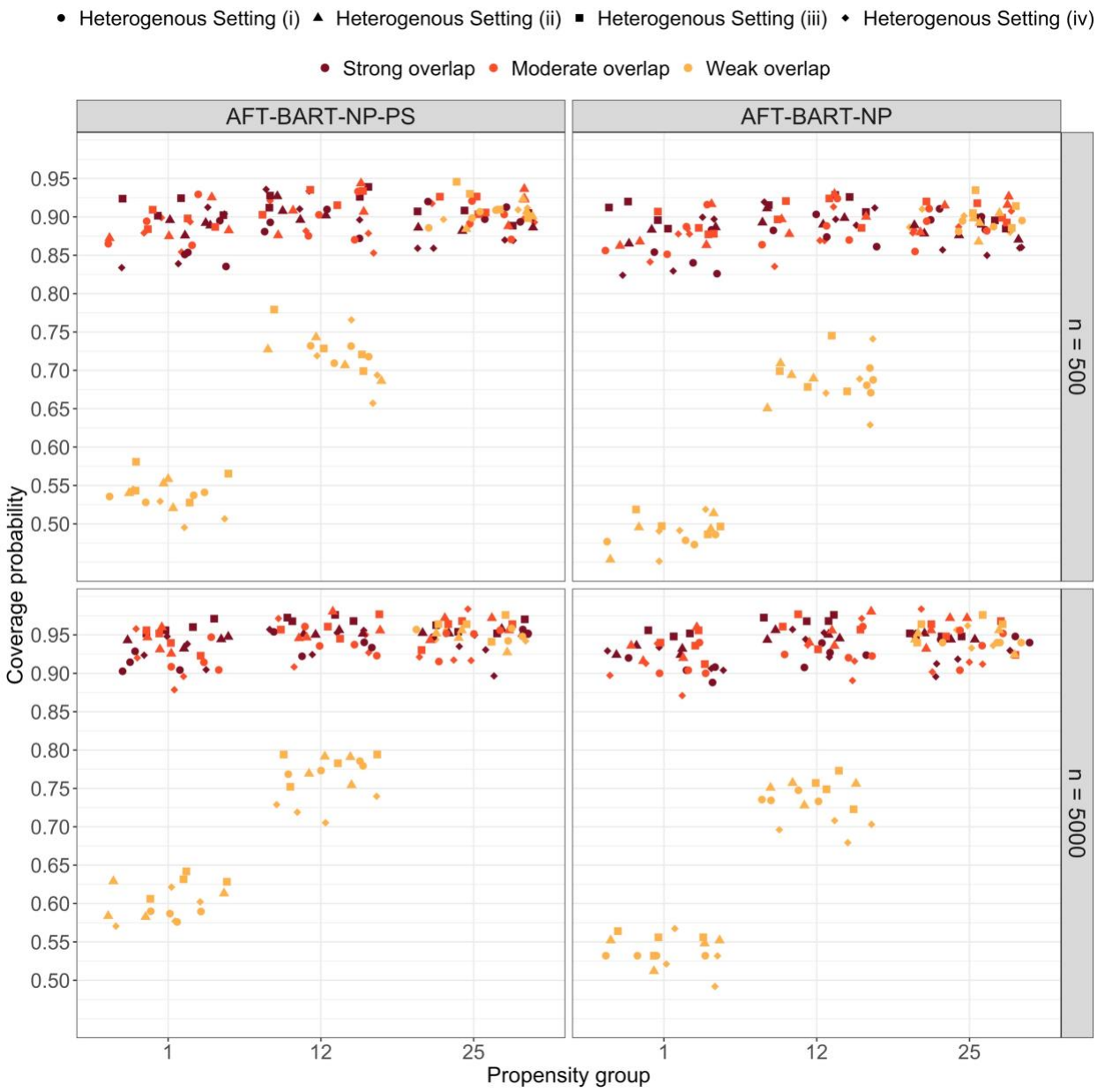
Web Figure 5: Relative biases among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 60% censoring. The relative bias for subgroup $k = 1, \ldots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



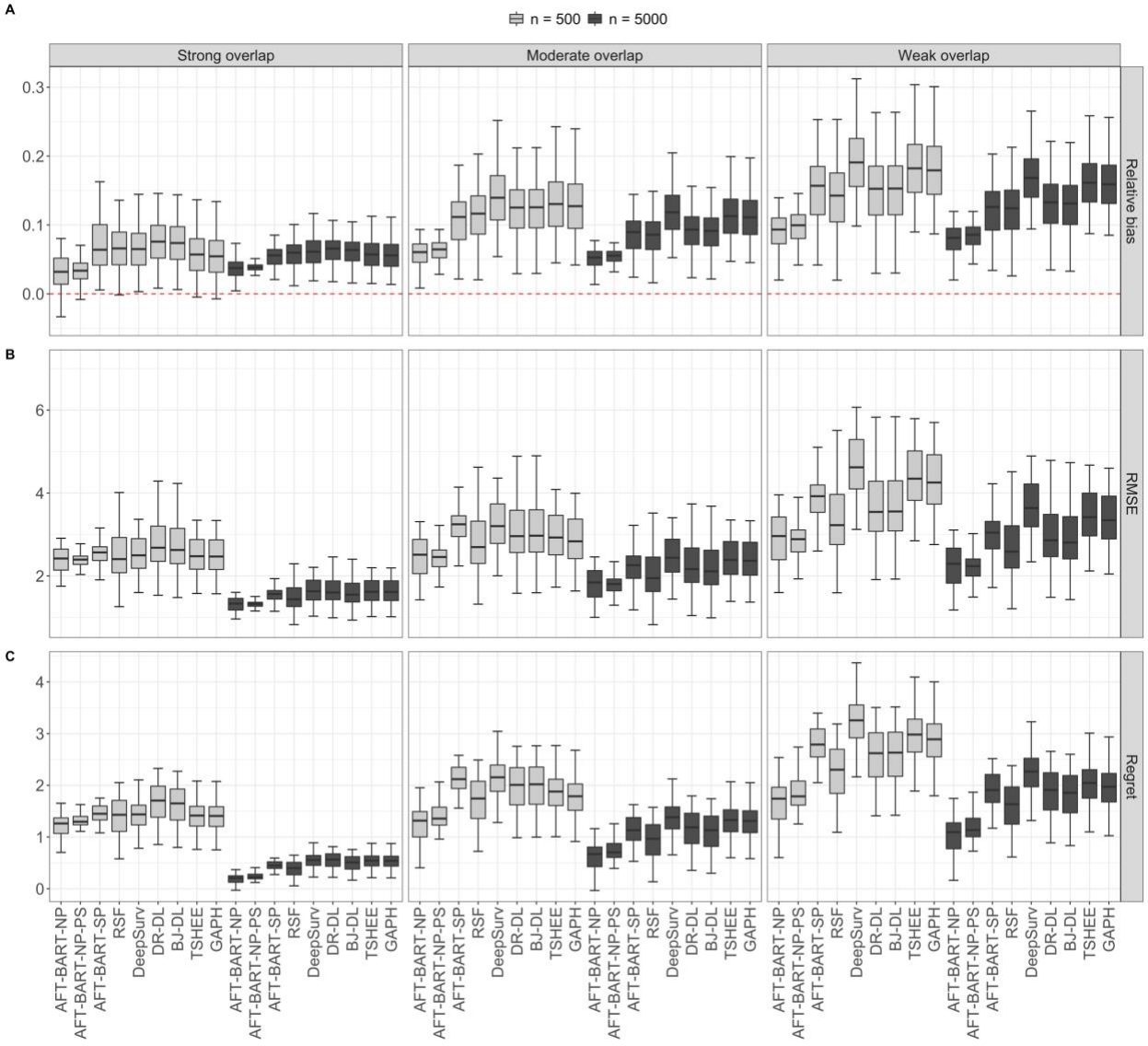
Web Figure 6: RMSE results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 60% censoring. The RMSE for subgroup $k = 1, \ldots, 50$ is defined by the root mean squared error between the estimated and true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



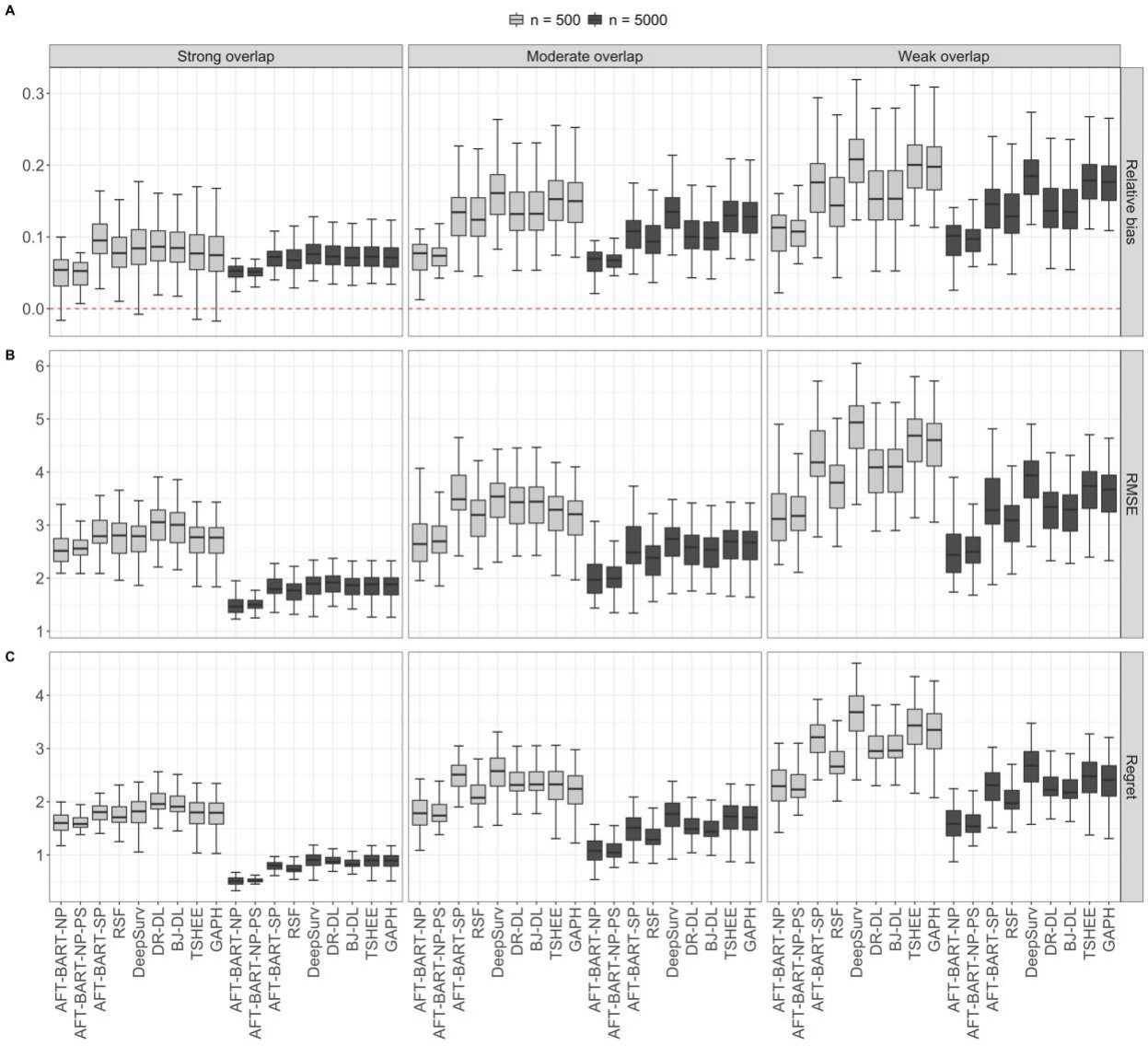
Web Figure 7: The regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong covariate overlap and 60% censoring. The RMSE for subgroup $k = 1, \ldots, 50$ is defined by the root mean squared error between the estimated and true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



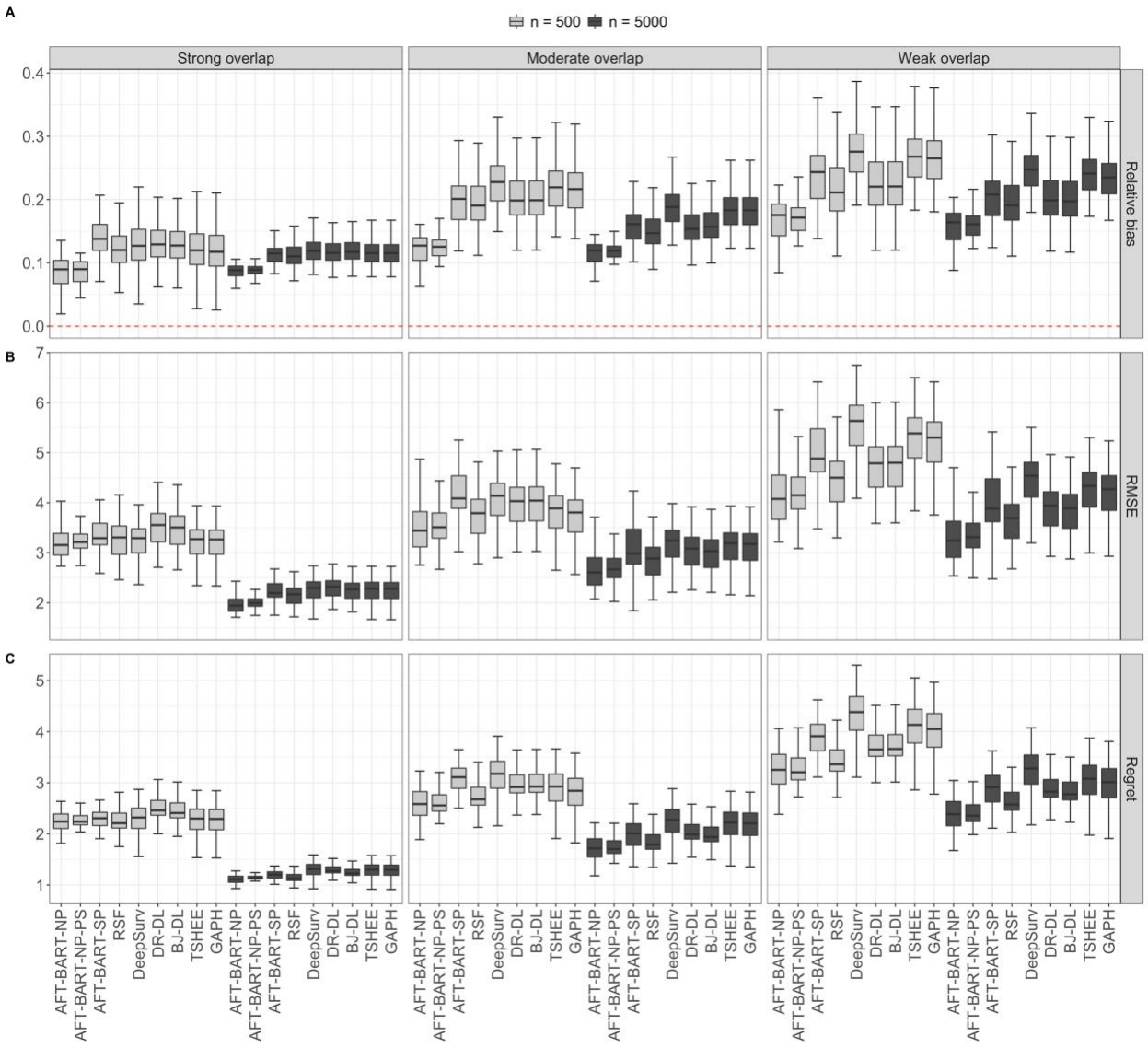
Web Figure 8. Dot plots of the coverage probability obtained from AFT-BART-NP-PS and AFT-BART-NP for 3 subgroups corresponding to extremely unbalanced (1), moderately balanced (12) and balanced (25) treatment assignment. Each colored cluster corresponds to a combination of simulation configurations representing PH vs. nPH, 20% censoring vs 60% censoring and heterogeneous setting (i), (ii) , (iii) versus (iv), for a given level of covariate overlap and a sample size.



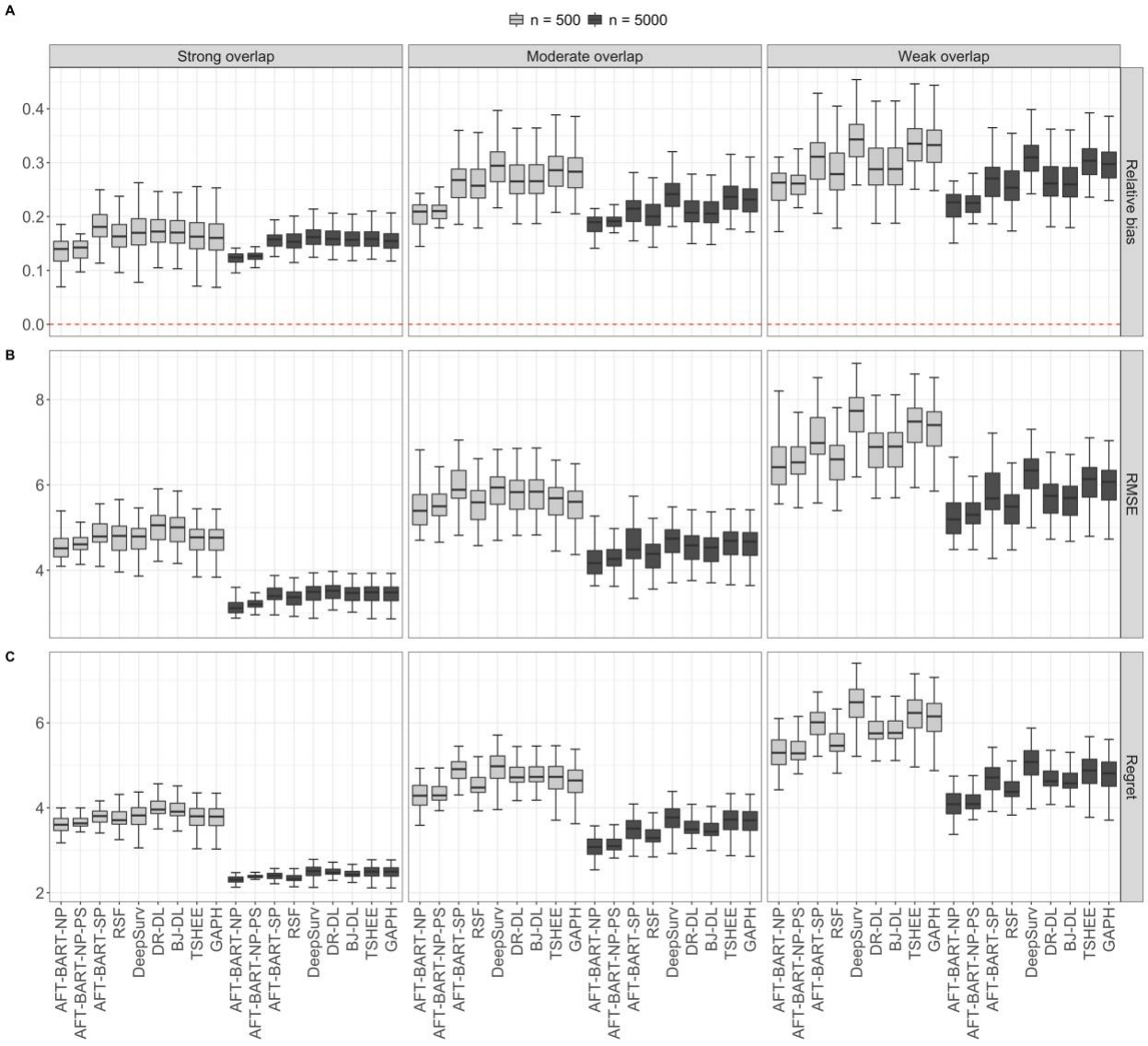
Web Figure 9. Relative biases, RMSE and regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong, moderate and weak covariate overlap, 60% covariate-dependent censoring, non-proportional hazards and heterogeneous setting (iv). The relative bias for subgroup $k = 1, \ldots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



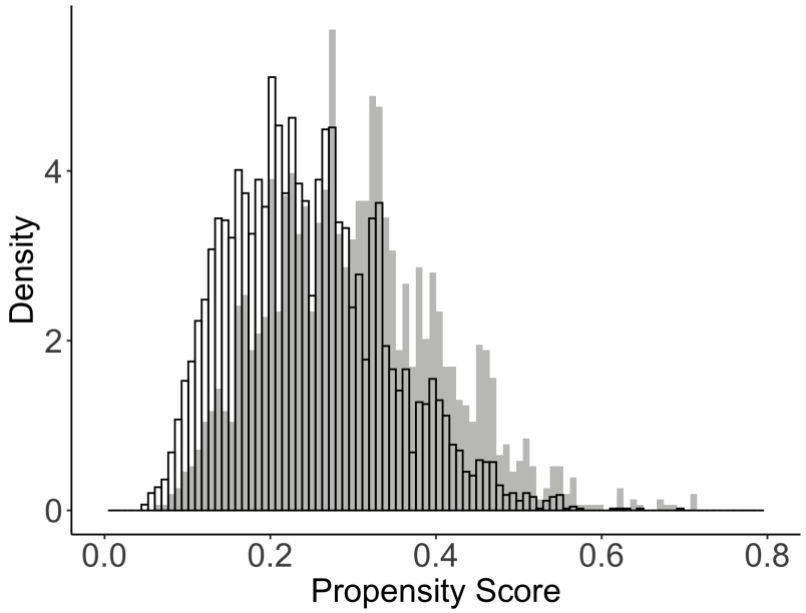
Web Figure 10. Sensitivity analysis results if we had no access to X_3. Relative biases, RMSE and regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong, moderate and weak covariate overlap, 60% covariate-dependent censoring, non-proportional hazards and heterogeneous setting (iii). The relative bias for subgroup $k = 1, \ldots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



Web Figure 11. Sensitivity analysis results if we had no access to X_3 and X_5. Relative biases, RMSE and regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong, moderate and weak covariate overlap, 60% covariate-dependent censoring, non-proportional hazards and heterogeneous setting (iii). The relative bias for subgroup $k = 1, \dots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



Web Figure 12. Sensitivity analysis results if we had no access to X_3, X_5 and X_6. Relative biases, RMSE and regret results among subgroups defined by quantiles of true propensity scores, for simulation scenarios with strong, moderate and weak covariate overlap, 60% covariate-dependent censoring, non-proportional hazards and heterogeneous setting (iii). The relative bias for subgroup $k = 1, \dots, 50$ is defined by the ratio of absolute bias and the true subgroup average survival causal effect. The boxplots are generated for 50 subgroups and across B simulation runs ($B = 250$ for $n = 5000$ and $B = 1000$ for $n = 500$).



Web Figure 13. Distribution of the propensity scores estimated via SuperLearner for two radiotherapy groups in the NCDB data. The un-shaded bars indicate EBRT plus AD (EBRT+AD) and the gray shaded bars indicate EBRT plus brachytherapy with or without AD (EBRT+brachy±AD). EBRT = external beam radiotherapy. AD = androgen deprivation.

Web Table 1. Mean (and standard deviation) of PEHE for each of 12 methods and each of 8 simulation configurations for $n = 500$ with strong overlap. CR: censoring rate; HS: heterogeneity setting; AFT-L: AFT-Lognormal; AFT-W: AFT-Weibull; ABN: AFT-BART-NP; ABNPS:AFT-BART-NP-PS; ABS:AFT-BART-SP.

		Proportional hazards				Non-proportional hazards			
CR		HS(i)	HS (ii)	HS (iii)	HS (iv)	HS(i)	HS (ii)	HS (iii)	HS (iv)
20%	AFT-L	3.63 (0.15)	4.48 (0.18)	4.41 (0.17)	4.89 (0.20)	4.08 (0.16)	4.65 (0.18)	4.71 (0.19)	5.10 (0.22)
	AFT-W	2.41 (0.13)	2.89 (0.13)	2.96 (0.13)	3.41 (0.18)	2.66 (0.14)	3.11 (0.15)	3.12 (0.15)	3.44 (0.17)
	Cox PH	2.48 (0.12)	2.93 (0.13)	2.91 (0.13)	3.39 (0.18)	4.11 (0.17)	4.68 (0.18)	4.78 (0.15)	5.12 (0.21)
	ABN	1.01 (0.18)	0.71 (0.17)	0.51 (0.15)	0.63 (0.16)	1.02 (0.18)	0.91 (0.16)	0.81 (0.15)	0.92 (0.16)
	ABNPS	1.02 (0.18)	0.72 (0.17)	0.52 (0.15)	0.64 (0.16)	1.03 (0.18)	0.92 (0.16)	0.82 (0.15)	0.93 (0.16)
	ABS	1.04 (0.18)	0.74 (0.17)	0.55 (0.15)	0.66 (0.16)	1.09 (0.18)	0.94 (0.16)	0.85 (0.15)	0.95 (0.16)
	RSFs	1.03 (0.18)	0.72 (0.17)	0.52 (0.16)	0.65 (0.16)	1.04 (0.18)	0.92 (0.16)	0.83 (0.15)	0.96 (0.16)
	DeepSu	1.09 (0.19)	0.81 (0.18)	0.61 (0.16)	0.72 (0.17)	1.11 (0.19)	0.98 (0.17)	0.87 (0.16)	1.00 (0.17)
	DR-DL	1.04 (0.18)	0.74 (0.17)	0.53 (0.15)	0.65 (0.16)	1.06 (0.18)	0.95 (0.16)	0.85 (0.15)	0.95 (0.16)

	BJ-DL	1.04 (0.18)	0.74 (0.17)	0.52 (0.15)	0.66 (0.16)	1.06 (0.18)	0.94 (0.16)	0.84 (0.15)	0.96 (0.16)
	TSHEE	1.08 (0.19)	0.77 (0.17)	0.58 (0.16)	0.68 (0.16)	1.09 (0.19)	0.96 (0.17)	0.86 (0.16)	0.97 (0.16)
	GAPH	1.08 (0.18)	0.79 (0.17)	0.57 (0.16)	0.78 (0.16)	1.10 (0.18)	0.99 (0.16)	0.86 (0.16)	1.07 (0.16)
		PH				nPH			
CR		HS(i)	HS (ii)	HS (iii)	HS (iv)	HS(i)	HS (ii)	HS (iii)	HS (iv)
60%	AFT-L	4.42 (0.18)	4.88 (0.18)	4.73 (0.19)	5.13 (0.20)	4.69 (0.19)	5.19 (0.21)	5.21 (0.21)	5.64 (0.23)
	AFT-W	2.93 (0.14)	3.53 (0.15)	3.41 (0.16)	4.01 (0.18)	3.18 (0.16)	4.01 (0.16)	4.03 (0.15)	4.47 (0.19)
	Cox PH	2.98 (0.14)	3.51 (0.15)	3.43 (0.15)	4.04 (0.18)	4.71 (0.19)	5.13 (0.21)	5.18 (0.21)	5.53 (0.23)
	ABN	1.31 (0.19)	0.82 (0.17)	0.72 (0.15)	0.82 (0.16)	1.49 (0.2)	1.01 (0.18)	0.91 (0.16)	1.01 (0.17)
	ABNPS	1.32 (0.19)	0.83 (0.17)	0.73 (0.15)	0.83 (0.16)	1.50 (0.2)	1.02 (0.18)	0.92 (0.16)	1.02 (0.17)
	ABS	1.35 (0.19)	0.85 (0.17)	0.75 (0.15)	0.84 (0.16)	1.54 (0.2)	1.06 (0.18)	0.95 (0.16)	1.05 (0.17)
	RSFs	1.32 (0.19)	0.83 (0.17)	0.73 (0.15)	0.83 (0.16)	1.51 (0.2)	1.02 (0.18)	0.92 (0.16)	1.02 (0.17)
	DeepSu	1.42 (0.2)	0.89 (0.18)	0.81 (0.16)	0.90 (0.17)	1.59 (0.21)	1.08 (0.19)	1.01 (0.17)	1.10 (0.18)
	DR-DL	1.35 (0.19)	0.85 (0.17)	0.74 (0.15)	0.84 (0.16)	1.54 (0.2)	1.06 (0.18)	0.94 (0.16)	1.03 (0.17)
	BJ-DL	1.35 (0.19)	0.84 (0.17)	0.73 (0.15)	0.85 (0.16)	1.53 (0.2)	1.05 (0.18)	0.93 (0.16)	1.06 (0.17)
	TSHEE	1.38 (0.19)	0.88 (0.18)	0.80 (0.16)	0.89 (0.17)	1.57 (0.21)	1.07 (0.19)	0.98 (0.17)	1.08 (0.17)
	GAPH	1.40 (0.19)	0.91 (0.17)	0.80 (0.16)	0.99 (0.17)	1.59 (0.2)	1.11 (0.18)	0.98 (0.17)	1.18 (0.17)

Web Table 2. Summary of frequentist coverage probability of the Bayesian credible intervals from the AFT-BART-NP for 5 propensity score subclasses, for each of two sample sizes, non-proportional hazards assumptions, and under varying degrees of overlap, in the scenario of heterogeneous setting (iv) and covariate-dependent censoring with 60% censoring proportion. The subclasses = 50 include units with the most extreme propensity scores, with the propensity scores closest to zero in G_k= 1 and the propensity scores closest to one in G_k= 50. The numbers in each cell represent the mean coverage of the ISTE for the corresponding subclass.

Sample size	G_k	Non-proportional hazards						Proportional hazards					
		Strong overlap		Moderate overlap		Weak overlap		Strong overlap		Moderate overlap		Weak overlap	
		X	*X+PS*	*X*	*X+PS*	*X*	*X+PS*	*X*	*X+PS*	*X*	*X+PS*	*X*	*X+PS*
$n = 5000$	1	0.92	0.93	0.90	0.90	0.49	0.53	0.91	0.93	0.90	0.91	0.48	0.54
	12	0.91	0.93	0.90	0.90	0.70	0.74	0.91	0.91	0.91	0.94	0.67	0.76
	25	0.93	0.94	0.91	0.92	0.91	0.92	0.92	0.92	0.91	0.93	0.91	0.93
	37	0.93	0.92	0.90	0.91	0.70	0.73	0.91	0.94	0.90	0.93	0.66	0.72
	50	0.94	0.94	0.90	0.90	0.48	0.55	0.90	0.93	0.89	0.91	0.50	0.55
$n = 500$	1	0.92	0.93	0.85	0.90	0.50	0.61	0.91	0.92	0.89	0.91	0.51	0.57
	12	0.90	0.93	0.90	0.88	0.70	0.71	0.91	0.93	0.90	0.91	0.65	0.73
	25	0.90	0.94	0.90	0.91	0.90	0.91	0.93	0.92	0.91	0.92	0.90	0.92
	37	0.90	0.93	0.88	0.90	0.70	0.75	0.91	0.93	0.91	0.92	0.65	0.71
	50	0.90	0.92	0.85	0.86	0.50	0.56	0.92	0.94	0.86	0.87	0.50	0.55

Web Table 3. Baseline characteristics of patients treated with EBRT plus AD versus EBRT plus brachytherapy with or without AD. EBRT = external beam radiotherapy. AD = androgen deprivation. SD = standard deviation.

	Overall (n = 7330)	EBRT+AD (n = 5430)	EBRT+brachy±AD (n = 1900)	p-value
Race (%)				0.122
White	5746 (78.4)	4245 (78.2)	1501 (79.0)	
Black	1327 (18.1)	1005 (18.5)	322 (16.9)	
Other	257 (3.5)	180 (3.3)	77 (4.1)	
Spanish or Hispanic Origin (%)				0.684
Non-Spanish; non-Hispanic	6997 (95.5)	5187 (95.5)	1810 (95.3)	
Spanish or Hispanic	333 (4.5)	243 (4.5)	90 (4.7)	
Insurance (%)				0.003
No	152 (2.1)	129 (2.4)	23 (1.2)	
Yes	7178 (97.9)	5301 (97.6)	1879 (98.8)	
Income (%)				<0.001
Less than $30,000	956 (13.0)	710 (13.1)	246 (12.9)	
$30,000 – $34,999	1304 (17.8)	979 (18.0)	325 (17.1)	
$35,000 – $45,999	2026 (27.6)	1586 (29.2)	440 (23.2)	
>$46,000	3044 (41.5)	2155 (39.7)	889 (46.8)	
Education† (%)				<0.001
<14%	2692 (36.7)	1894 (34.9)	798 (42.0)	
14%-19.9%	1801 (24.6)	1368 (25.2)	433 (22.8)	
20% - 28.9%	1728 (23.6)	1336 (24.6)	392 (20.6)	
≥29%	1109 (15.1)	832 (15.3)	277 (14.6)	
Charlson comorbidity index (%)				0.41
0	6187 (84.4)	4576 (84.3)	1611 (84.8)	
1‡	925 (12.6)	684 (12.6)	241 (12.7)	
>1§	218 (3.0)	170 (3.1)	48 (2.5)	
Clinical T Stage (%)				0.267
≤cT2	6103 (83.3)	4505 (83.0)	1688 (84.1)	
≥ cT3	1227 (16.7)	925 (17.0)	302 (15.9)	
Year of Diagnosis (%)				<0.001
2004-2007	54 (0.7)	25 (0.5)	29 (1.5)	
2008-2010	1728 (23.6)	1132 (20.8)	596 (31.4)	
2010-2013	5548 (75.7)	4273 (78.7)	1275 (67.1)	
Age (mean [SD])	68.9 (8.2)	69.5 (8.2)	67.2 (7.8)	<0.001
PSA (ng/mL) (mean [SD])	20.9 (22.7)	21.7 (23.4)	18.5 (20.4)	<0.001
Gleason score				<0.001
6	256 (3.5)	163 (3.0)	93 (4.9)	
7	1282 (17.5)	930 (17.1)	352 (18.5)	
8	3170 (43.2)	2336 (43.0)	834 (43.9)	
9	2388 (32.6)	1816 (33.4)	572 (30.1)	
10	234 (3.2)	185 (3.4)	49 (2.6)	

†Percentage of adults in the patient's zip code who did not graduate from high school.

‡Myocardial Infarction, Congestive Heart Failure, Peripheral Vascular Disease, Cerebrovascular Disease, Dementia, Chronic Pulmonary Disease, Rheumatologic Disease, Peptic Ulcer Disease, Mild Liver Disease, Diabetes

§Diabetes with Chronic Complications, Hemiplegia or Paraplegia, Renal Disease

Web Table 4. Understanding heterogeneous survival effects (based on median survival) of EBRT plus brachytherapy with or without AD (EBRT+brachy±AD) versus EBRT plus AD (EBRT+AD) using linear regression with dependent variable being the MCMC samples of ISTE from the AFT-BART-NP-PS model. EBRT = external beam radiotherapy. AD = androgen deprivation.

	Estimate	95% Credible Interval
Intercept	5.80	(–3.12, 7.03)
Race (reference = White)		
Black	0.11	(–0.10, 0.37)
Other	0.37	(–0.11, 0.91)
Spanish or Hispanic Origin (Yes vs. No)	0.23	(–0.30, 0.70)
Insurance (Yes vs. No)	–0.49	(–1.07, 0.26)
Education† (reference = <14%)		
14% - 19.9%	0.11	(–0.10, 0.39)
20% - 28.9%	0.23	(–0.05, 0.45)
≥29%	0.28	(–0.02, 0.56)
Charlson comorbidity index (%)		
1‡ vs. 0	–0.03	(–0.34, 0.20)
>1§ vs. 0	–0.36	(–0.90, 0.28)
Clinical T Stage (%)		
≤cT2 vs. ≥ cT3	–0.15	(–0.39, 0.08)
Year of Diagnosis (%)		
2008-2010 vs. 2004-2007	–0.68	(–1.72, 0.22)
2010-2013 vs. 2004-2007	–0.66	(–1.57, 0.35)
Age (mean [SD])	–0.69	(–0.79, –0.59)
PSA (ng/mL) (mean [SD])	3.99	(3.60, 4.38)
Gleason score		
7 vs. 6	–0.25	(–0.85, 0.25)
8 vs. 6	–0.34	(–0.90, 0.20)
9 vs. 6	–0.43	(–1.00, 0.10)
10 vs. 6	–0.53	(–1.07, 0.09)
Income (reference = less than $30,000)		
$30,000 – $34,999	0.16	(–0.20, 0.50)
$35,000 – $45,999	–0.04	(–0.34, 0.26)
>$46,000	–0.17	(–0.43, 0.13)

†Percentage of adults in the patient's zip code who did not graduate from high school.
‡Myocardial Infarction, Congestive Heart Failure, Peripheral Vascular Disease, Cerebrovascular Disease, Dementia, Chronic Pulmonary Disease, Rheumatologic Disease, Peptic Ulcer Disease, Mild Liver Disease, Diabetes
§Diabetes with Chronic Complications, Hemiplegia or Paraplegia, Renal Disease