WEB-BASED SUPPLEMENTARY MATERIALS FOR A BAYESIAN CREDIBLE SUBGROUPS APPROACH TO IDENTIFYING PATIENT SUBGROUPS WITH POSITIVE TREATMENT EFFECTS

PATRICK M. SCHNELL[1,*], QI TANG[2], WALTER W. OFFEN[2], AND BRADLEY P. CARLIN[1]

[1] DIVISION OF BIOSTATISTICS, UNIVERSITY OF MINNESOTA SCHOOL OF PUBLIC HEALTH, MINNEAPOLIS, MINNESOTA, U.S.A.

[2] ABBVIE, NORTH CHICAGO, ILLINOIS, U.S.A.

These supplementary materials tabulate additional simulation study results and display results of a sensitivity analysis of the real data example. Tables 1–3 present credible subgroup coverage and model fit statistics, and Tables 4–6 present sensitivities and specificities of D and S when treated as a diagnostic test for benefit. The situations simulated are six parameter values of the linear model with γ_1 the baseline treatment effect, γ_2 the effect of a binary covariate, and γ_3 the effect of a continuous covariate uniformly distributed on $[-3, 3]$. Each situation and sample size has 1000 trials simulated and credible subgroups are computed at the 80% level. BART fits are performed with the default arguments in the `R` package `BayesTree` and 500 posterior samples kept after 100 burn-in samples.

Table 1 presents the credible subgroup coverage and model fit statistics for simulations with a sample size of $n = 40$, which reflects the sample size of the example dataset ($n = 41$) presented in the paper. The total coverage measures the frequency with which $\mathrm{D} \subseteq \mathrm{B}_{\boldsymbol{\gamma}} \subseteq \mathrm{S}$. The restricted covariate space (RCS) and highest posterior density (HPD) methods are conservative with respect to total coverage when the linear model is correctly specified, while the pure Bayes (PB), pointwise (PW), and BART methods are sometimes conservative and sometimes anticonservative. However, the conservative methods have larger pair sizes, which measure the proportion of covariate points in the uncertainty region $\mathrm{S} \setminus \mathrm{D}$.

Effect MSE is the mean squared error of the personalized treatment effect for each covariate point, and measures model fit. The same linear model is fit for each of the credible subgroup methods excluding BART. The linear model outperforms BART when there is treatment effect heterogeneity with respect to the continuous covariate (excepting the inverted U scenario), otherwise BART outperforms the linear model. This reflects the tendency of BART to split along binary covariates and the fact that estimates are constant within each cell. An F test of the null hypothesis $\gamma_2 = \gamma_3 = 0$ is also carried out as a test for treatment

Date: *Received Month* YYYY. *Revised Month* YYYY. *Accepted Month* YYYY..

effect heterogeneity. This test generally performs well, except in the inverted U scenario, in which case there is no linear trend in the treatment effect even though there is heterogeneity.

Table 2 shows that the above properties also hold with an increased sample size of $n = 100$, though the pair sizes and effect MSEs generally decrease. BART, in addition to having a better model fit in the cases in which it did for $n = 40$, has a lower treatment effect MSE for the linear case $\boldsymbol{\gamma} = (1, 1, 1)$. The linear model retains its advantage in other cases with continuous treatment effect heterogeneity, including the misspecified square root and S-curve scenarios. Table 3 shows that when the sample size reaches $n = 350$, the BART model matches or outperforms the linear model with respect to treatment effect MSE except in the correctly specified $\boldsymbol{\gamma} = (0, 0, 1)$ and $\boldsymbol{\gamma} = (0, 1, 1)$ cases.

Tables 4–6 show that, as expected, the methods that generate credible subgroups with more conservative coverage have a lower sensitivity of D and specificity of S while having a higher specificity of D and sensitivity of S.

Figure 1 shows the credible subgroup plots resulting from using varying levels of informativeness in the prior for treatment-covariate interactions. Using a less informative prior variance of 100 (rather than 1) for standardized treatment-covariate interactions yielded similar results to those presented in the primary analysis of the example data. Using a more informative prior variance of 1/100 shrank the interaction estimates to near zero, resulting in a much more homogeneous treatment effect estimate across the covariate space, and a larger exclusive credible subgroup.

Table 1. Coverage and model fit statistics for 80% credible subgroup pairs (n=40).

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Effect MSE	Heterog. Tests
$\gamma = (0,0,0)$	PB	0.46	0.46	1.00	0.75	0.24	0.18
	RCS	0.88	0.88	1.00	0.95	0.24	0.18
	HPD	0.91	0.91	1.00	0.97	0.24	0.18
	PW	0.43	0.43	1.00	0.59	0.24	0.18
	BART	0.68	0.68	1.00	0.71	0.10	–
$\gamma = (0,0,1)$	PB	0.82	0.87	0.94	0.25	0.24	1.00
	RCS	0.94	0.95	0.99	0.34	0.24	1.00
	HPD	0.96	0.97	0.99	0.38	0.24	1.00
	PW	0.46	0.67	0.77	0.13	0.24	1.00
	BART	0.59	0.78	0.81	0.45	1.62	–
$\gamma = (0,1,0)$	PB	0.55	0.55	0.99	0.55	0.28	0.45
	RCS	0.87	0.87	1.00	0.78	0.28	0.45
	HPD	0.91	0.91	1.00	0.82	0.28	0.45
	PW	0.47	0.47	0.97	0.39	0.28	0.45
	BART	0.49	0.49	1.00	0.40	0.21	–
$\gamma = (0,1,1)$	PB	0.77	0.84	0.93	0.25	0.28	1.00
	RCS	0.92	0.93	0.98	0.35	0.28	1.00
	HPD	0.95	0.96	0.99	0.38	0.28	1.00
	PW	0.41	0.61	0.76	0.14	0.28	1.00
	BART	0.48	0.54	0.94	0.38	1.63	–
$\gamma = (1,0,0)$	PB	0.99	1.00	0.99	0.25	1.21	0.18
	RCS	1.00	1.00	1.00	0.50	1.21	0.18
	HPD	1.00	1.00	1.00	0.56	1.21	0.18
	PW	0.97	1.00	0.97	0.13	1.21	0.18
	BART	1.00	1.00	1.00	0.04	0.13	–
$\gamma = (1,1,1)$	PB	0.73	0.79	0.94	0.24	1.53	1.00
	RCS	0.92	0.93	0.99	0.33	1.53	1.00
	HPD	0.94	0.95	1.00	0.35	1.53	1.00
	PW	0.43	0.62	0.80	0.15	1.53	1.00
	BART	0.12	0.12	1.00	0.14	1.59	–
Square Root	PB	0.64	0.86	0.77	0.62	0.28	0.56
	RCS	0.92	0.97	0.95	0.84	0.28	0.56
	HPD	0.94	0.98	0.97	0.87	0.28	0.56
	PW	0.38	0.74	0.61	0.42	0.28	0.56
	BART	0.54	0.83	0.71	0.67	0.35	–
S-curve	PB	0.76	0.85	0.90	0.44	0.35	0.92
	RCS	0.93	0.95	0.98	0.61	0.35	0.92
	HPD	0.95	0.96	0.99	0.65	0.35	0.92
	PW	0.42	0.65	0.74	0.24	0.35	0.92
	BART	0.57	0.76	0.80	0.60	0.87	–
Inverted U	PB	0.20	0.41	0.70	0.73	0.37	0.18
	RCS	0.80	0.82	0.96	0.93	0.37	0.18
	HPD	0.85	0.86	0.98	0.95	0.37	0.18
	PW	0.16	0.30	0.60	0.56	0.37	0.18
	BART	0.42	0.55	0.86	0.66	0.19	–

Note: Total coverage rates at or above 80% and low pair sizes (analogous to interval lengths for interval estimation) are desired. The heterogeneity test is a Bayesian F test for the null hypothesis $\gamma_2 = \gamma_3 = 0$, and is not performed on BART fits.

Table 2. Coverage and model fit statistics for 80% credible subgroup pairs (n=100).

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Effect MSE	Heterog. Tests
$\boldsymbol{\gamma} = (0,0,0)$	PB	0.45	0.45	1.00	0.75	0.11	0.18
	RCS	0.88	0.88	1.00	0.95	0.11	0.18
	HPD	0.92	0.92	1.00	0.97	0.11	0.18
	PW	0.43	0.43	1.00	0.59	0.11	0.18
	BART	0.57	0.57	1.00	0.69	0.07	–
$\boldsymbol{\gamma} = (0,0,1)$	PB	0.79	0.85	0.93	0.14	0.11	1.00
	RCS	0.94	0.95	0.99	0.19	0.11	1.00
	HPD	0.96	0.97	0.99	0.21	0.11	1.00
	PW	0.45	0.64	0.77	0.08	0.11	1.00
	BART	0.61	0.77	0.83	0.15	0.48	–
$\boldsymbol{\gamma} = (0,1,0)$	PB	0.53	0.53	1.00	0.41	0.13	0.77
	RCS	0.89	0.89	1.00	0.58	0.13	0.77
	HPD	0.92	0.92	1.00	0.62	0.13	0.77
	PW	0.53	0.53	1.00	0.31	0.13	0.77
	BART	0.48	0.48	1.00	0.32	0.11	–
$\boldsymbol{\gamma} = (0,1,1)$	PB	0.76	0.83	0.92	0.14	0.12	1.00
	RCS	0.92	0.94	0.98	0.20	0.12	1.00
	HPD	0.95	0.96	0.99	0.22	0.12	1.00
	PW	0.45	0.63	0.76	0.08	0.12	1.00
	BART	0.56	0.67	0.88	0.17	0.47	–
$\boldsymbol{\gamma} = (1,0,0)$	PB	1.00	1.00	1.00	0.04	1.12	0.18
	RCS	1.00	1.00	1.00	0.17	1.12	0.18
	HPD	1.00	1.00	1.00	0.21	1.12	0.18
	PW	1.00	1.00	1.00	0.02	1.12	0.18
	BART	1.00	1.00	1.00	0.01	0.08	–
$\boldsymbol{\gamma} = (1,1,1)$	PB	0.76	0.84	0.92	0.16	1.26	1.00
	RCS	0.93	0.95	0.98	0.22	1.26	1.00
	HPD	0.96	0.96	0.99	0.24	1.26	1.00
	PW	0.44	0.64	0.77	0.09	1.26	1.00
	BART	0.36	0.38	0.96	0.17	0.45	–
Square Root	PB	0.68	0.92	0.76	0.50	0.15	0.89
	RCS	0.90	0.98	0.92	0.68	0.15	0.89
	HPD	0.93	0.99	0.94	0.72	0.15	0.89
	PW	0.37	0.77	0.55	0.29	0.15	0.89
	BART	0.52	0.84	0.67	0.53	0.20	–
S-curve	PB	0.78	0.87	0.91	0.26	0.22	1.00
	RCS	0.93	0.95	0.98	0.35	0.22	1.00
	HPD	0.96	0.97	0.99	0.39	0.22	1.00
	PW	0.42	0.65	0.72	0.14	0.22	1.00
	BART	0.60	0.78	0.81	0.24	0.36	–
Inverted U	PB	0.20	0.34	0.77	0.70	0.24	0.18
	RCS	0.77	0.79	0.97	0.92	0.24	0.18
	HPD	0.83	0.84	0.98	0.94	0.24	0.18
	PW	0.14	0.24	0.68	0.53	0.24	0.18
	BART	0.28	0.42	0.81	0.60	0.14	–

Note: Total coverage rates at or above 80% and low pair sizes (analogous to interval lengths for interval estimation) are desired. The heterogeneity test is a Bayesian F test for the null hypothesis $\gamma_2 = \gamma_3 = 0$, and is not performed on BART fits.

TABLE 3. Coverage and model fit statistics for 80% credible subgroup pairs (n=350).

Truth	Method	Total Coverage	D Coverage	S Coverage	Pair Size	Effect MSE	Heterog. Tests
$\boldsymbol{\gamma} = (0,0,0)$	PB	0.43	0.43	1.00	0.76	0.04	0.21
	RCS	0.88	0.88	1.00	0.95	0.04	0.21
	HPD	0.91	0.91	1.00	0.97	0.04	0.21
	PW	0.41	0.41	1.00	0.59	0.04	0.21
	BART	0.39	0.39	1.00	0.72	0.03	–
$\boldsymbol{\gamma} = (0,0,1)$	PB	0.79	0.84	0.95	0.07	0.04	1.00
	RCS	0.96	0.96	0.99	0.10	0.04	1.00
	HPD	0.97	0.97	0.99	0.11	0.04	1.00
	PW	0.56	0.66	0.87	0.04	0.04	1.00
	BART	0.70	0.80	0.90	0.09	0.11	–
$\boldsymbol{\gamma} = (0,1,0)$	PB	0.48	0.48	1.00	0.34	0.04	1.00
	RCS	0.92	0.92	1.00	0.48	0.04	1.00
	HPD	0.93	0.93	1.00	0.48	0.04	1.00
	PW	0.54	0.54	1.00	0.29	0.04	1.00
	BART	0.45	0.45	1.00	0.35	0.04	–
$\boldsymbol{\gamma} = (0,1,1)$	PB	0.77	0.82	0.95	0.07	0.04	1.00
	RCS	0.95	0.96	0.99	0.10	0.04	1.00
	HPD	0.96	0.97	0.99	0.11	0.04	1.00
	PW	0.54	0.64	0.87	0.04	0.04	1.00
	BART	0.65	0.75	0.88	0.09	0.10	–
$\boldsymbol{\gamma} = (1,0,0)$	PB	1.00	1.00	1.00	0.00	1.04	0.21
	RCS	1.00	1.00	1.00	0.00	1.04	0.21
	HPD	1.00	1.00	1.00	0.00	1.04	0.21
	PW	1.00	1.00	1.00	0.00	1.04	0.21
	BART	1.00	1.00	1.00	0.00	0.03	–
$\boldsymbol{\gamma} = (1,1,1)$	PB	0.77	0.82	0.94	0.08	1.09	1.00
	RCS	0.93	0.94	0.99	0.12	1.09	1.00
	HPD	0.96	0.96	1.00	0.13	1.09	1.00
	PW	0.52	0.65	0.87	0.05	1.09	1.00
	BART	0.63	0.71	0.90	0.11	0.10	–
Square Root	PB	0.72	0.97	0.74	0.26	0.07	1.00
	RCS	0.87	0.99	0.87	0.36	0.07	1.00
	HPD	0.92	1.00	0.92	0.39	0.07	1.00
	PW	0.37	0.88	0.43	0.14	0.07	1.00
	BART	0.61	0.82	0.78	0.26	0.07	–
S-curve	PB	0.79	0.86	0.93	0.12	0.14	1.00
	RCS	0.96	0.96	0.99	0.17	0.14	1.00
	HPD	0.97	0.98	0.99	0.18	0.14	1.00
	PW	0.50	0.66	0.80	0.07	0.14	1.00
	BART	0.76	0.80	0.95	0.07	0.08	–
Inverted U	PB	0.15	0.18	0.90	0.61	0.16	0.22
	RCS	0.61	0.62	0.98	0.86	0.16	0.22
	HPD	0.70	0.70	0.99	0.89	0.16	0.22
	PW	0.08	0.10	0.82	0.43	0.16	0.22
	BART	0.30	0.46	0.74	0.47	0.06	–

Note: Total coverage rates at or above 80% and low pair sizes (analogous to interval lengths for interval estimation) are desired. The heterogeneity test is a Bayesian F test for the null hypothesis $\gamma_2 = \gamma_3 = 0$, and is not performed on BART fits.

Table 4. Diagnostic properties of 80% credible subgroup pairs (n=40).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\boldsymbol{\gamma} = (0,0,0)$	PB	–	0.87	–	0.12
	RCS	–	0.97	–	0.02
	HPD	–	0.98	–	0.01
	PW	–	0.79	–	0.20
	BART	–	0.85	–	0.13
$\boldsymbol{\gamma} = (0,0,1)$	PB	0.76	0.99	1.00	0.74
	RCS	0.67	1.00	1.00	0.64
	HPD	0.64	1.00	1.00	0.61
	PW	0.87	0.98	0.99	0.84
	BART	0.54	0.97	0.98	0.51
$\boldsymbol{\gamma} = (0,1,0)$	PB	0.68	0.83	1.00	0.05
	RCS	0.38	0.95	1.00	0.01
	HPD	0.33	0.96	1.00	0.00
	PW	0.79	0.71	0.99	0.13
	BART	0.79	0.62	1.00	0.03
$\boldsymbol{\gamma} = (0,1,1)$	PB	0.81	0.99	1.00	0.64
	RCS	0.75	1.00	1.00	0.52
	HPD	0.72	1.00	1.00	0.48
	PW	0.89	0.97	0.99	0.78
	BART	0.78	0.88	1.00	0.29
$\boldsymbol{\gamma} = (1,0,0)$	PB	0.75	–	1.00	–
	RCS	0.50	–	1.00	–
	HPD	0.44	–	1.00	–
	PW	0.87	–	1.00	–
	BART	0.96	–	1.00	–
$\boldsymbol{\gamma} = (1,1,1)$	PB	0.87	0.97	1.00	0.39
	RCS	0.82	0.99	1.00	0.25
	HPD	0.80	0.99	1.00	0.21
	PW	0.92	0.93	0.99	0.59
	BART	0.97	0.47	1.00	0.03
Square Root	PB	0.28	0.98	0.96	0.41
	RCS	0.13	1.00	1.00	0.19
	HPD	0.10	1.00	1.00	0.15
	PW	0.45	0.95	0.92	0.58
	BART	0.17	0.93	0.88	0.30
S-curve	PB	0.56	0.99	0.99	0.54
	RCS	0.40	1.00	1.00	0.38
	HPD	0.35	1.00	1.00	0.34
	PW	0.74	0.97	0.98	0.72
	BART	0.36	0.94	0.96	0.34
Inverted U	PB	0.21	0.81	0.94	0.07
	RCS	0.06	0.95	0.99	0.01
	HPD	0.04	0.96	1.00	0.01
	PW	0.33	0.71	0.89	0.14
	BART	0.30	0.75	0.95	0.06

Note: Statistics are averaged without undefined values, e.g. sensitivity of D when $B_{\boldsymbol{\gamma}}$ is empty.

TABLE 5. Diagnostic properties of 80% credible subgroup pairs (n=100).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\boldsymbol{\gamma} = (0, 0, 0)$	PB	–	0.88	–	0.13
	RCS	–	0.98	–	0.02
	HPD	–	0.98	–	0.02
	PW	–	0.80	–	0.21
	BART	–	0.85	–	0.16
$\boldsymbol{\gamma} = (0, 0, 1)$	PB	0.87	1.00	1.00	0.85
	RCS	0.82	1.00	1.00	0.79
	HPD	0.81	1.00	1.00	0.78
	PW	0.92	0.99	0.99	0.90
	BART	0.85	0.99	0.99	0.83
$\boldsymbol{\gamma} = (0, 1, 0)$	PB	0.93	0.83	1.00	0.08
	RCS	0.78	0.96	1.00	0.01
	HPD	0.73	0.97	1.00	0.01
	PW	0.96	0.73	1.00	0.16
	BART	0.97	0.68	1.00	0.06
$\boldsymbol{\gamma} = (0, 1, 1)$	PB	0.89	0.99	1.00	0.80
	RCS	0.85	1.00	1.00	0.73
	HPD	0.84	1.00	1.00	0.71
	PW	0.93	0.98	0.99	0.87
	BART	0.90	0.98	1.00	0.72
$\boldsymbol{\gamma} = (1, 0, 0)$	PB	0.96	–	1.00	–
	RCS	0.83	–	1.00	–
	HPD	0.79	–	1.00	–
	PW	0.98	–	1.00	–
	BART	0.99	–	1.00	–
$\boldsymbol{\gamma} = (1, 1, 1)$	PB	0.91	0.99	1.00	0.61
	RCS	0.88	1.00	1.00	0.49
	HPD	0.87	1.00	1.00	0.45
	PW	0.95	0.96	0.99	0.74
	BART	0.95	0.85	1.00	0.33
Square Root	PB	0.39	0.99	0.98	0.58
	RCS	0.24	1.00	0.99	0.40
	HPD	0.20	1.00	1.00	0.35
	PW	0.60	0.98	0.94	0.75
	BART	0.31	0.98	0.93	0.53
S-curve	PB	0.75	0.99	1.00	0.73
	RCS	0.66	1.00	1.00	0.64
	HPD	0.62	1.00	1.00	0.60
	PW	0.86	0.98	0.98	0.83
	BART	0.76	0.99	0.99	0.73
Inverted U	PB	0.26	0.79	0.95	0.06
	RCS	0.08	0.94	1.00	0.01
	HPD	0.06	0.96	1.00	0.00
	PW	0.39	0.67	0.91	0.11
	BART	0.40	0.76	0.96	0.08

Note: Statistics are averaged without undefined values, e.g. sensitivity of D when $B_{\boldsymbol{\gamma}}$ is empty.

TABLE 6. Diagnostic properties of 80% credible subgroup pairs (n=350).

Truth	Method	Sensitivity of D	Specificity of D	Sensitivity of S	Specificity of S
$\gamma = (0, 0, 0)$	PB	–	0.88	–	0.12
	RCS	–	0.98	–	0.02
	HPD	–	0.98	–	0.02
	PW	–	0.80	–	0.21
	BART	–	0.86	–	0.14
$\gamma = (0, 0, 1)$	PB	0.94	1.00	1.00	0.92
	RCS	0.91	1.00	1.00	0.89
	HPD	0.91	1.00	1.00	0.88
	PW	0.96	0.99	1.00	0.94
	BART	0.92	0.99	1.00	0.89
$\gamma = (0, 1, 0)$	PB	1.00	0.81	1.00	0.12
	RCS	1.00	0.97	1.00	0.02
	HPD	1.00	0.98	1.00	0.01
	PW	1.00	0.76	1.00	0.17
	BART	1.00	0.79	1.00	0.09
$\gamma = (0, 1, 1)$	PB	0.95	1.00	1.00	0.90
	RCS	0.93	1.00	1.00	0.86
	HPD	0.92	1.00	1.00	0.85
	PW	0.97	0.99	1.00	0.93
	BART	0.93	0.99	1.00	0.87
$\gamma = (1, 0, 0)$	PB	1.00	–	1.00	–
	RCS	1.00	–	1.00	–
	HPD	1.00	–	1.00	–
	PW	1.00	–	1.00	–
	BART	1.00	–	1.00	–
$\gamma = (1, 1, 1)$	PB	0.96	0.99	1.00	0.80
	RCS	0.94	1.00	1.00	0.72
	HPD	0.93	1.00	1.00	0.70
	PW	0.97	0.98	1.00	0.86
	BART	0.95	0.98	1.00	0.70
Square Root	PB	0.63	1.00	0.99	0.83
	RCS	0.53	1.00	0.99	0.75
	HPD	0.49	1.00	1.00	0.72
	PW	0.76	1.00	0.96	0.91
	BART	0.65	0.99	0.98	0.79
S-curve	PB	0.89	1.00	1.00	0.86
	RCS	0.84	1.00	1.00	0.82
	HPD	0.83	1.00	1.00	0.81
	PW	0.93	0.99	0.99	0.91
	BART	0.94	0.99	1.00	0.92
Inverted U	PB	0.39	0.69	0.99	0.02
	RCS	0.15	0.89	1.00	0.00
	HPD	0.12	0.92	1.00	0.00
	PW	0.55	0.55	0.96	0.06
	BART	0.61	0.87	0.98	0.17

Note: Statistics are averaged without undefined values, e.g. sensitivity of D when B_{γ} is empty.

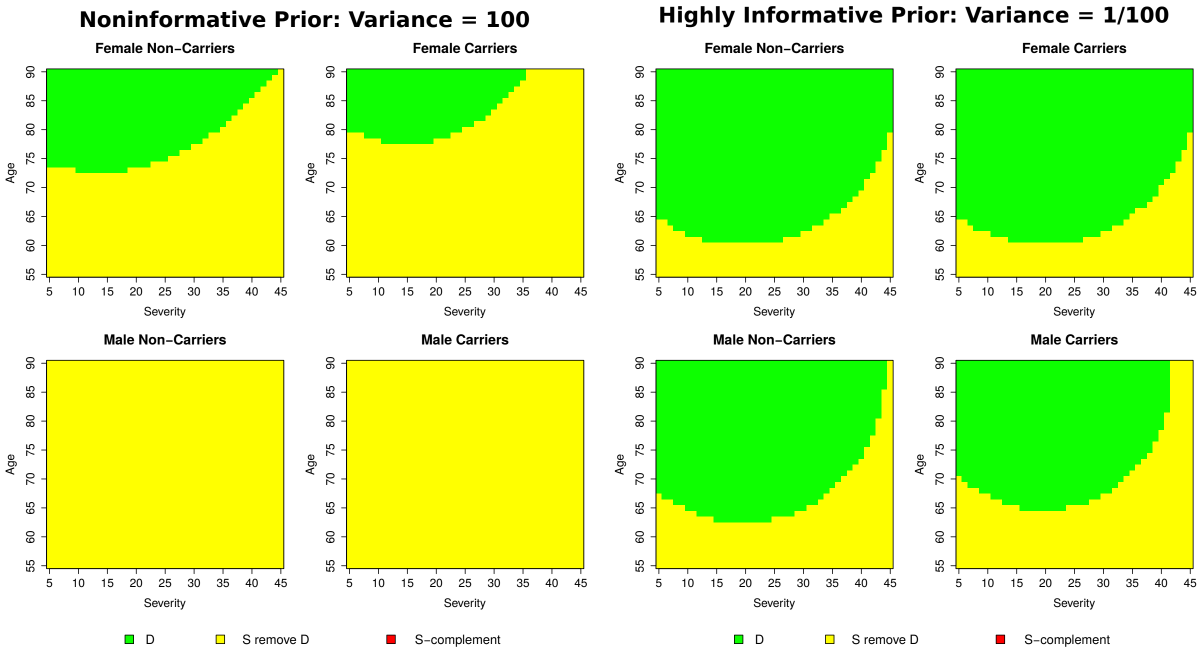

FIGURE 1. Reducing the informativeness of the prior for treatment-covariate interactions in the example analysis (left) has a very small effect on the resulting 80% credible subgroups, but strongly increasing the informativeness (right) dramatically expands the exclusive credible subgroup, approaching the homogeneous result that would be obtained by not modeling interactions.

E-mail address: SCHN0956@UMN.EDU