Supplementary material: A modified Random Survival Forests algorithm for non-recurring, time to event outcomes ascertained using imperfect, self-reports or laboratory based diagnostic tests

Hui Xu[1], Xiangdong Gu[1], Mahlet G. Tadesse[2], Raji Balasubramanian[1]

[1]Department of Biostatistics and Epidemiology

University of Massachusetts Amherst, Amherst, MA 01003

[2] Department of Mathematics and Statistics

Georgetown University, Washington, DC 20057

email: rbalasub@umass.edu

1 Organization

The material in this supplement is organized as follows: In Section 1, we present simulation results to illustrate the degradation in the performance of the original RSF algorithm in the presence of error-prone outcomes. In Section 2, we present results from simulations based on data from a cardiovascular disease Omics study. In Section 3, we present additional results from the application of the proposed algorithm to GWAS data from subjects in the WHI.

2 Effects of error in self-reported outcomes on variable selection by Random Survival Forests (RSF)

We illustrate the degradation in the variable selection performance of the original RSF algorithm, with increasing error in the self-reported outcomes.

Each simulated dataset included $N = 100$ subjects and $P = 100$ covariates, of which the first five $(Z_1, \cdots, Z_5)$ were assumed to be true biomarkers. We assumed that the duration of follow-up was 4 years and that there were annual visits at which self-reported outcomes were collected. We assumed that there were no missed visits. Each covariate was simulated according to an independent standard normal distribution. We assumed that the true time to event followed an exponential distribution and that the set of five biomarkers influenced the

outcome through a proportional hazards model. Let λ_0 denote the hazard corresponding to the reference group (corresponding to $Z_1 = \cdots = Z_5 = 0$). Under the proportional hazards model, the hazard for a subject with arbitrary values of the covariates $Z_1, \cdots, Z_5$ is given by:

$$\lambda_{Z_1,Z_2,Z_3,Z_4,Z_5} \quad = \quad \lambda_0 e^{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5}$$

We set the values of λ_0 to correspond to values of cumulative incidence for the reference group of 0.15, 0.25 and 0.5, respectively. The values of $\beta_1, \cdots, \beta_5$ were set to 2. For each subject i, observed values of the binary, self-reported outcomes at visits at years 1-4 ($R_{i1}, \cdots, R_{i4}$) were simulated by assuming specific values for the sensitivity and specificity of self-reports.

For example, assume that the simulated time-to-event for subject i is $X_i = 2.5$ years, the sensitivity and specificity of self-reported outcomes are $\varphi_1 = 0.9$ and $\varphi_0 = 0.7$, respectively. Then, the self-reported outcomes at visits 1-4 are simulated according to $\mathrm{P}(R_{i1} = 1 \mid X_i = 2.5, t_{i1} = 1) = \mathrm{P}(R_{i2} = 1 \mid X_i = 2.5, t_{i2} = 2) = 1 - \varphi_0$ and $\mathrm{P}(R_{i3} = 1 \mid X_i = 2.5, t_{i3} = 3) = \mathrm{P}(R_{i4} = 1 \mid X_i = 2.5, t_{i4} = 4) = \varphi_1$.

We fit the RSF algorithm to each simulated dataset, by setting the number of trees to 1000. The variables ranking among the top 5% were considered as "discovered biomarkers". Averaging over 100 simulations for each setting, we estimated the proportion of datasets in

which each of the true five biomarkers ($Z_1 = \cdots = Z_5$) was "discovered".

Panels (a) - (b) of Figure 1 present the proportion of times that true biomarkers were ranked among the top 5 by RSF as a function of sensitivity or specificity. We consider different parameter settings corresponding to (sensitivity, specificity) of the self-reported outcome and cumulative incidence (in the reference group). The results show that the reduction in specificity has a deleterious effect on variable selection, when sensitivity is assumed to be perfect (Figure 1 of Supplement, Panel (a)). On the other hand, when specificity is fixed at 1, reduction in sensitivity has a modest effect on variable selection (Figure 1, Panel (b)). When sensitivity, specificity and cumulative incidence were set to 0.61, 0.995 and 0.15 respectively (characteristics of diabetes self-reports in the WHI), we observed that the true five biomarkers were discovered among the top five variables on average 66% of the time - in comparison, when self-reports are perfect, the true five biomarkers were discovered among the top five variables on average 80% of the time.

3 Cardiovascular disease omics Study

The cardiovascular disease 'omics' omics study that was conducted to discover prognostic biomarkers in blood plasma for near-term cardiovascular events. Subjects were selected from

the CATHGEN project, which collected peripheral blood samples from consenting research subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 through 2011. 68 cases were selected from among individuals who had a major adverse cardiac event (MACE) within two years following the time of their sample collection. In a 1:1 matched study design, 68 controls were selected from individuals who were MACE-free for the two years following sample collection and were matched to cases on age, gender, race/ethnicity and severity of coronary artery disease. High-content mass spectrometry and multiplexed immunoassay-based techniques were employed to quantify 625 proteins and metabolites from each subject's serum specimen. Comprehensive metabolite profiling of the individual samples was based on a combination of four platforms employing mass spectrometry (MS) based techniques to profile lipids, fatty acids, amino acids, sugars and other metabolites. Proteomic analysis was based on a combination of targeted methods using a quantitative multiplexed immunoassay technique as well as a comprehensive protein profiling strategy based on tandem mass spectrometry. A detailed description of the mass spectrometry based platforms and proteomics analysis can be found in a previous publication (Guo and Balasubramanian, 2012).

3.1 Illustrating variable importance from a single tree

A single tree was trained on a bootstrap sample from a simulated dataset to illustrate the proposed variable importance metric. The covariate matrix was obtained as a random subset of 100 out the 625 covariates for all 136 subjects in the cardiovascular disease omics study. The event times were assumed to follow an exponential distribution, where the first five covariates (denoted $Z_1, \cdots, Z_5$) from the 'omics' study were assumed to be associated with the outcome through a Cox proportional hazards model. Let λ_0 denote the hazard corresponding to the reference group (corresponding to $Z_1 = \cdots = Z_5 = 0$). Under the proportional hazards model, the hazard for a subject with arbitrary values of the covariates $Z_1, \cdots, Z_5$ is given by $\lambda_{Z_1,Z_2,Z_3,Z_4,Z_5} = \lambda_0 e^{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5}$. The regression coefficients were set to $\beta_1 = \cdots = \beta_5 = 1.0$. The hazard function for the reference group λ_0 was set such that the cumulative incidence rate during the four year follow-up period was 0.1. We assumed that the duration of follow-up was 4 years, that there were annual visits at which self-reported outcomes were collected, with sensitivity and specificity of 1.0 and with no missed visits. For each subject i, binary self-reported outcomes at each visit at years 1-4 ($R_{i1}, \cdots, R_{i4}$) were simulated by assuming perfect sensitivity and specificity of self-reports.

A single tree was trained on a bootstrap sample of the simulated dataset, where each split was selected from among a random subset of 10 covariates. In Figure 2 of the Supplement, we show a realization of the fitted tree - of the five true biomarkers (labeled Var1-Var5),

Var3 was selected as the primary (first level) splitting variable and Var2 selected as a third level splitting variable. In Figure 3 [Panel (a)], the difference between the observed log likelihood for each subject i in the OOB dataset and the average log likelihood for each subject i over 100 permutations of a covariate that is associated with outcome in the Cox model is shown (that is, $l_i - \tilde{l}_{i.}$). In comparison, Figure 3, Panel (b), shows the same metric for each OOB subject i, averaged over 100 permutations of a noise covariate. Figure 3 shows that when a variable that is truly associated with the outcome is permuted, this results in a significant decrease in l_i (and a corresponding increase in variable importance) when compared to permutations of a noise covariate.

3.2 Simulation details

The simulation study described in Section 3.3 of the main paper incorporated the structure of observed data by setting the covariate matrix in each simulation to equal a randomly selected subset of 100 covariates for all 136 subjects in the cardiovascular disease omics study. Each of the 100 covariates was standardized to render it with mean 0 and unit variance. The distribution of the Pearson correlation between pairs of selected covariates is shown in Figure 4. The pairwise Pearson correlations range from -0.54 to 0.99. Each simulated dataset was generated according to the description in Section 3 of the main paper. Figure 5 shows the marginal distributions of the standardized values of each of the five selected biomarkers in

the Cox proportional hazards model . Table 1 presents the average proportions of simulated datasets in which the five true biomarkers were 'discovered', for the settings of (1) No missing data and (2) Missing all data following the first positive self-report, respectively.

4 Application: Women's Health Initiative Clinical Trials and Observational Study SHARe

Table 2 presents a summary of the baseline characteristics of the 9873 subjects in the WHI Clinical Trials and Observational Study SHARe. Table 3 presents a summary of genes that either contain or flank the SNPs identified among the top 10 by at least one analysis presented in this paper that have been previously reported in association with Type 2 diabetes. Figure 6 shows a bar plot of the variable importance of each SNP (1 through 88,277) resulting from the analysis based on the proposed algorithm - the horizontal dashed line indicates the variable importance threshold separating the top 10 SNPs from the rest.

REFERENCES

Bolton, K., Segal, D., McMillan, J., Jowett, J., Heilbronn, L., Abberton, K., Zimmet, P., Chisholm, D., Collier, G., and Walder, K. (2008). Decorin is a secreted protein associated with obesity and type 2 diabetes. *International journal of obesity*, 32(7):1113–1121.

Dong, C., Beecham, A., Slifer, S., Wang, L., McClendon, M. S., Blanton, S. H., Rundek, T., and Sacco, R. L. (2011). Genome-wide linkage and peak-wide association study of obesity-related quantitative traits in caribbean hispanics. *Human genetics*, 129(2):209–219.

Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J., and Mohlke, K. L. (2014). Identification of a regulatory variant that binds foxa1 and foxa2 at the cdc123/camk1d type 2 diabetes gwas locus. *PLoS Genet*, 10(9):e1004633.

Guo, Y. and Balasubramanian, R. (2012). Comparative evaluation of classifiers in the presence of statistical interaction between features in high-dimensionality data settings. *International Journal of Biostatistics*, 8.

McDonough, C. W., Palmer, N. D., Hicks, P. J., Roh, B. H., An, S. S., Cooke, J. N., Hester, J. M., Wing, M. R., Bostrom, M. A., Rudock, M. E., et al. (2011). A genome-wide association study for diabetic nephropathy genes in african americans. *Kidney international*, 79(5):563–572.

Murea, M., Lu, L., Ma, L., Hicks, P. J., Divers, J., McDonough, C. W., Langefeld, C. D., Bowden, D. W., and Freedman, B. I. (2011). Genome-wide association scan for survival on dialysis in african-americans with type 2 diabetes. *American journal of nephrology*, 33(6):502–509.

Palmer, N. D., McDonough, C. W., Hicks, P. J., Roh, B. H., Wing, M. R., An, S. S., Hester, J. M., Cooke, J. N., Bostrom, M. A., Rudock, M. E., et al. (2012). A genome-wide association search for type 2 diabetes genes in african americans. *PloS one*, 7(1):e29202.

Sakai, K., Imamura, M., Tanaka, Y., Iwata, M., Hirose, H., Kaku, K., Maegawa, H., Watada, H., Tobe, K., Kashiwagi, A., et al. (2013). Replication study for the association of 9 east asian gwas-derived loci with susceptibility to type 2 diabetes in a japanese population. *PloS one*, 8(9):e76317.

Yeh, S.-H., Chang, W.-C., Chuang, H., Huang, H.-C., Liu, R.-T., and Yang, K. D. (2016). Differentiation of type 2 diabetes mellitus with different complications by proteomic analysis of plasma low abundance proteins. *Journal of Diabetes & Metabolic Disorders*, 15(1):24.

			No missing data		Missing all data following first positive self-report	
$1 - \boldsymbol{S_{j+1}}$	$\boldsymbol{\varphi_1}$	$\boldsymbol{\varphi_0}$	$\boldsymbol{p_{RSF}}$	$\boldsymbol{p_1}$	$\boldsymbol{p_{RSF}}$	$\boldsymbol{p_1}$
0.10	1.00	1.00	0.692(±0.0462)	0.700(±0.0458)	0.692(±0.0462)	0.728(±0.0445)
	0.75	1.00	0.634(±0.0482)	0.722(±0.0448)	0.634(±0.0482)	0.736(±0.0441)
	0.61	0.995	0.574(±0.0494)	0.702(±0.0457)	0.574(±0.0494)	0.706(±0.0456)
	1.00	0.90	0.464(±0.0499)	0.698(±0.0459)	0.464(±0.0499)	0.534(±0.0499)
0.30	1.00	1.00	0.778(±0.0416)	0.752(±0.0432)	0.778(±0.0416)	0.772(±0.0420)
	0.75	1.00	0.742(±0.0438)	0.76(±0.0427)	0.742(±0.0438)	0.772(±0.0420)
	0.61	0.995	0.700(±0.0458)	0.744(±0.0436)	0.700(±0.0458)	0.738(±0.0438)
	1.00	0.90	0.654(±0.0476)	0.748(±0.0434)	0.654(±0.0476)	0.638(±0.0481)

Table 1: **Simulation - Cardiovascular Disease Omics Study**: The average proportion of datasets (±SE) in which the five true biomarkers are ranked among the top five according to three measures of variable importance, namely (1) original RSF algorithm (p_{RSF}); and (2) variable importance from the modified RSF algorithm (p_1). $1 - S_{J+1}, \varphi_1, \varphi_0$ denote the cumulative incidence in the reference group, sensitivity and specificity, respectively.

Gene Symbol	Association Type	PubMed Example
DCN	Biomarker	Bolton et al., 2008
WWOX	Genetic Variation	Sakai et al., 2013
RYR2*	Biomarker	Palmer et al., 2012; Dong et al., 2011
DACH1*	Biomarker	McDonough et al., 2011
ESRRG*	Biomarker, Genetic Variation	Murea et al., 2011
USH2A	Biomarker, Genetic Variation	Yeh et al., 2016
GPATCH2*	Biomarker, Genetic Variation	Murea et al., 2011
CDC123	Biomarker, Genetic Variation	Fogarty et al., 2014

Table 3: Genes selected by the modified RSF algorithm that were reported to be related to risk of type 2 diabetes in previous literature. * indicates genes that were reported to be associated with type 2 diabetes in African Americans or Hispanic Americans.

Continuous Variables	**Mean(SD)**
Age(years)	60.65(6.768)
Dietary Energy Intake	1121.56(932.210)
Minutes of recreational physical activity per week	153.41(171.475)
Body Mass Index	30.02(6.043)
Categorical Variables	**N** (%)
Smoking Status	
Never smoked	5372 (54%)
Past smoker	3567 (36%)
Current smoker	934 (10 %)
Alcohol Intake	
Non-drinker	1597 (16%)
Past drinker	2693 (27%)
<1 drink per month	1375 (14%)
<1 drink per week	2088 (21%)
1 to <7 drink per week	1647 (17%)
7+ drink per week	473 (5%)
Hormone Therapy Use	
Never use hormones	4310 (43 %)
Past hormone user	2340 (24%)
current hormone user	3223 (33%)
Family History of Diabetes	
No	4838 (49%)
Yes	4198 (43 %)
Don't know	837 (8 %)
Education	
<8 grade	1244 (13%)
High school	1404 (14%)
College	3743 (38%)
Post-graduation	3482 (35%)

Table 2: Baseline characteristics of the subjects in the WHI Clinical Trials and Observational Study SHARe ($n = 9,873$).

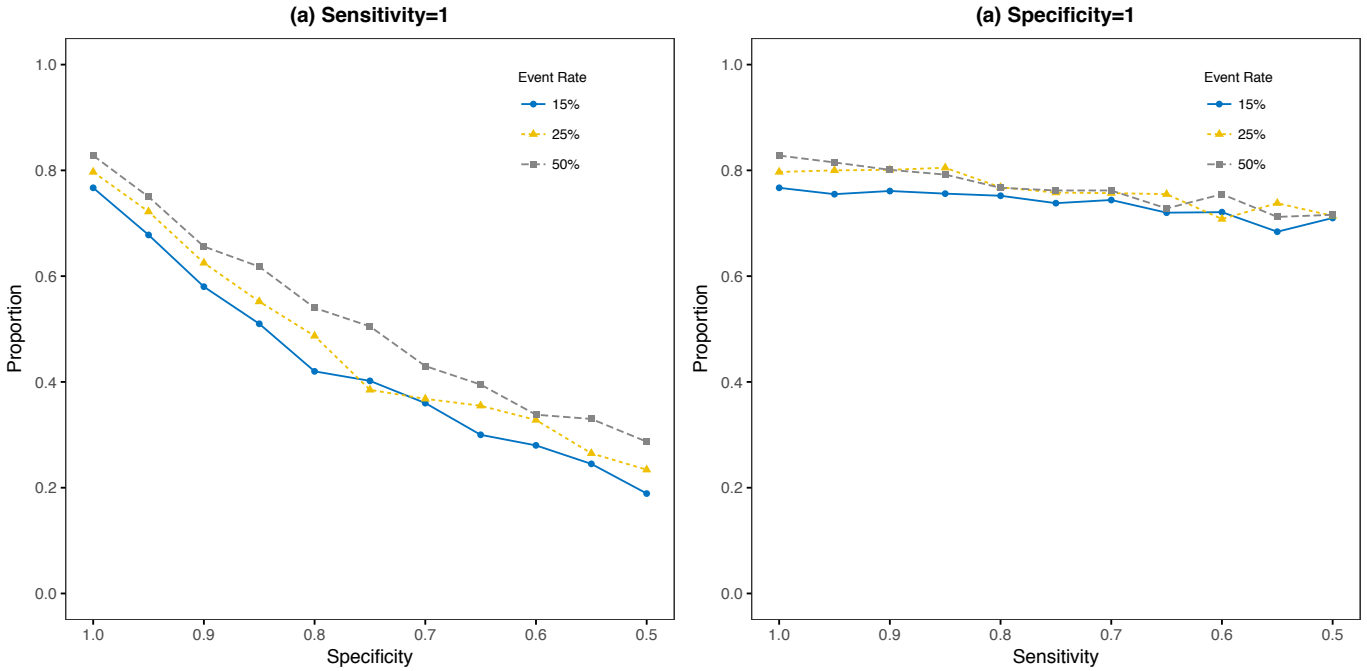

Figure 1: **Effects of error in self-reported outcomes on Random Survival Forests.** Proportion of datasets in which the true biomarkers ranked among the top five by Random Survival Forests, under different parameter settings with respect to sensitivity (φ_1), specificity (φ_0) and cumulative incidence in the reference group during study period. Data were simulated assuming no missed visits.

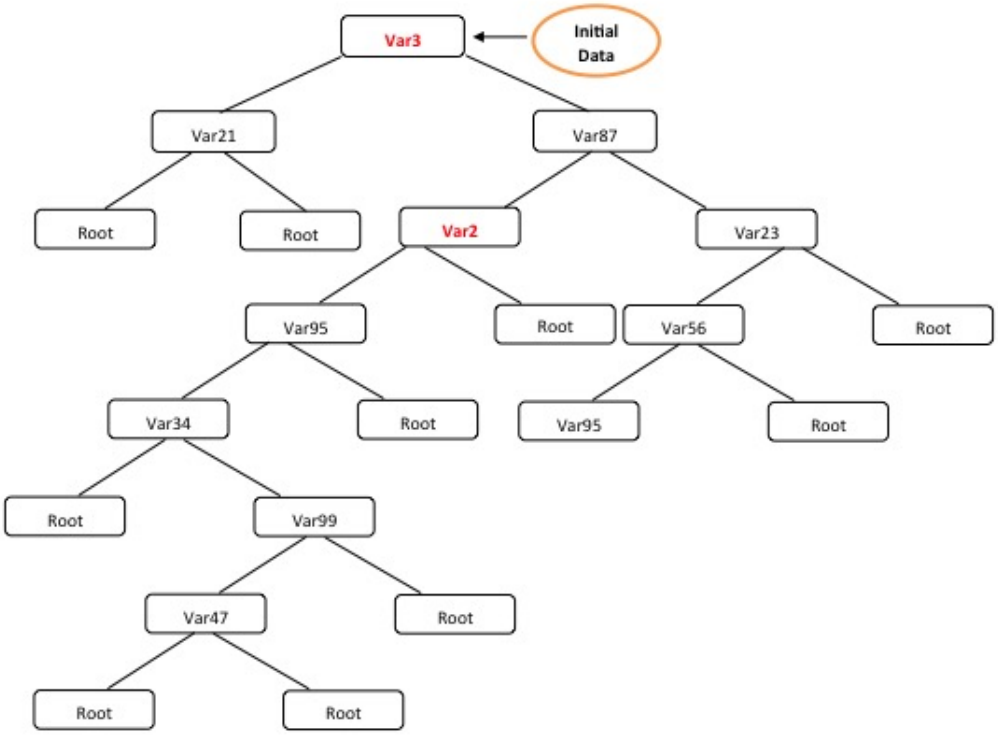

Figure 2: **Cardiovascular Disease Omics Study.** Realization of a single tree fit to an bootstrap dataset. Var1-Var5 (red) denote true biomarkers and Var6-Var100 denote covariates with no association with outcome.

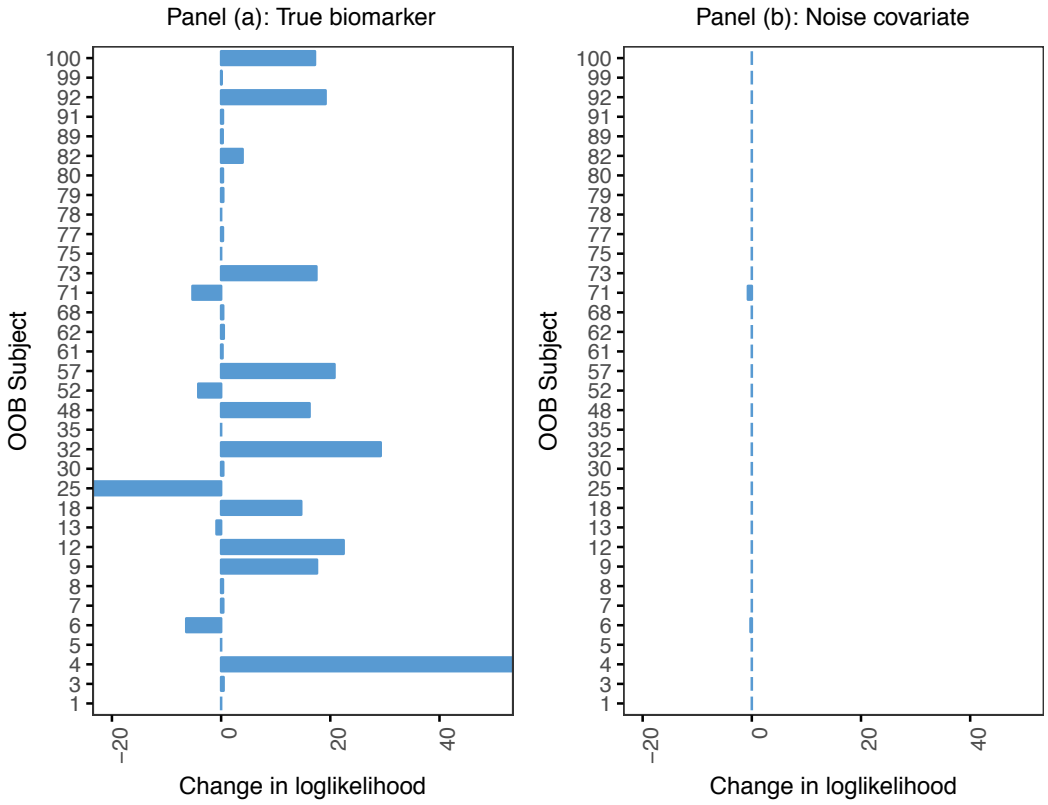

Figure 3: **Cardiovascular Disease Omics Study.** The average change in log likelihood for each subject i in the OOB dataset over $k = 1 \cdots 100$ permutations of a specific covariate is shown $(l_i - \tilde{l}_{i.})$. Panel (a) Permuted covariate is associated with the outcome in the Cox proportional hazards model (true biomarker); Panel (b) Permuted covariate is not associated with outcome (noise covariate).

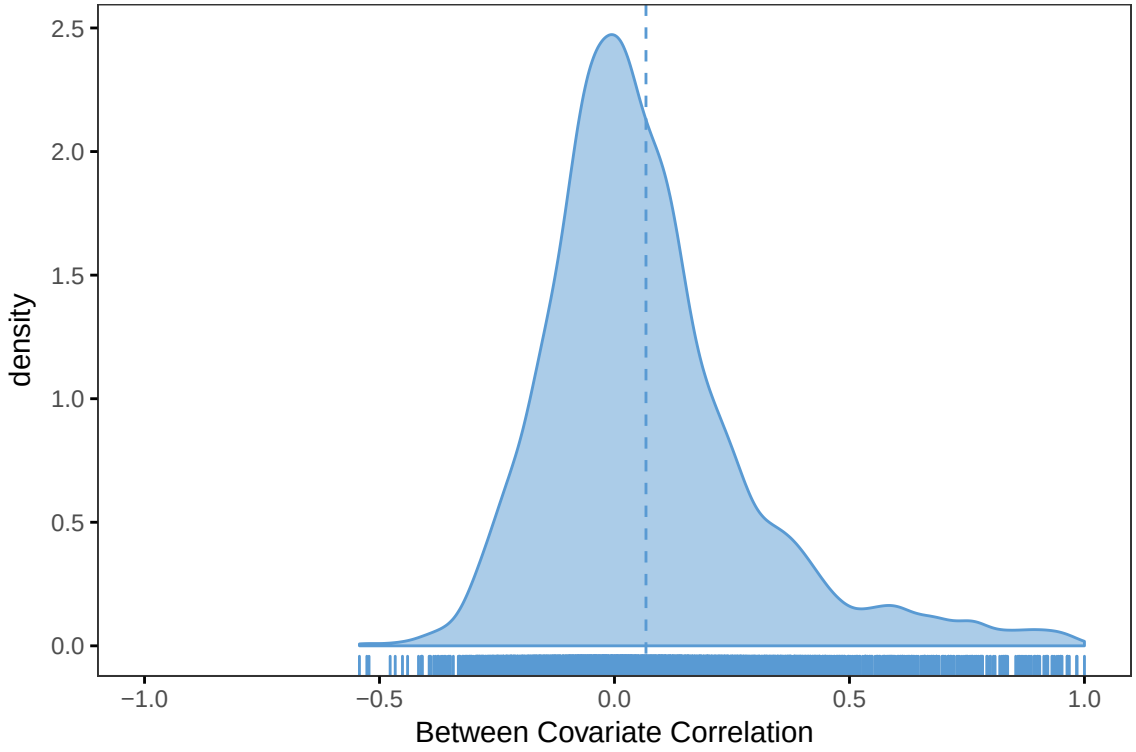

Figure 4: **Cardiovascular Disease Omics Study.** Distribution of Pearson correlation between pairs of 100 selected covariates.

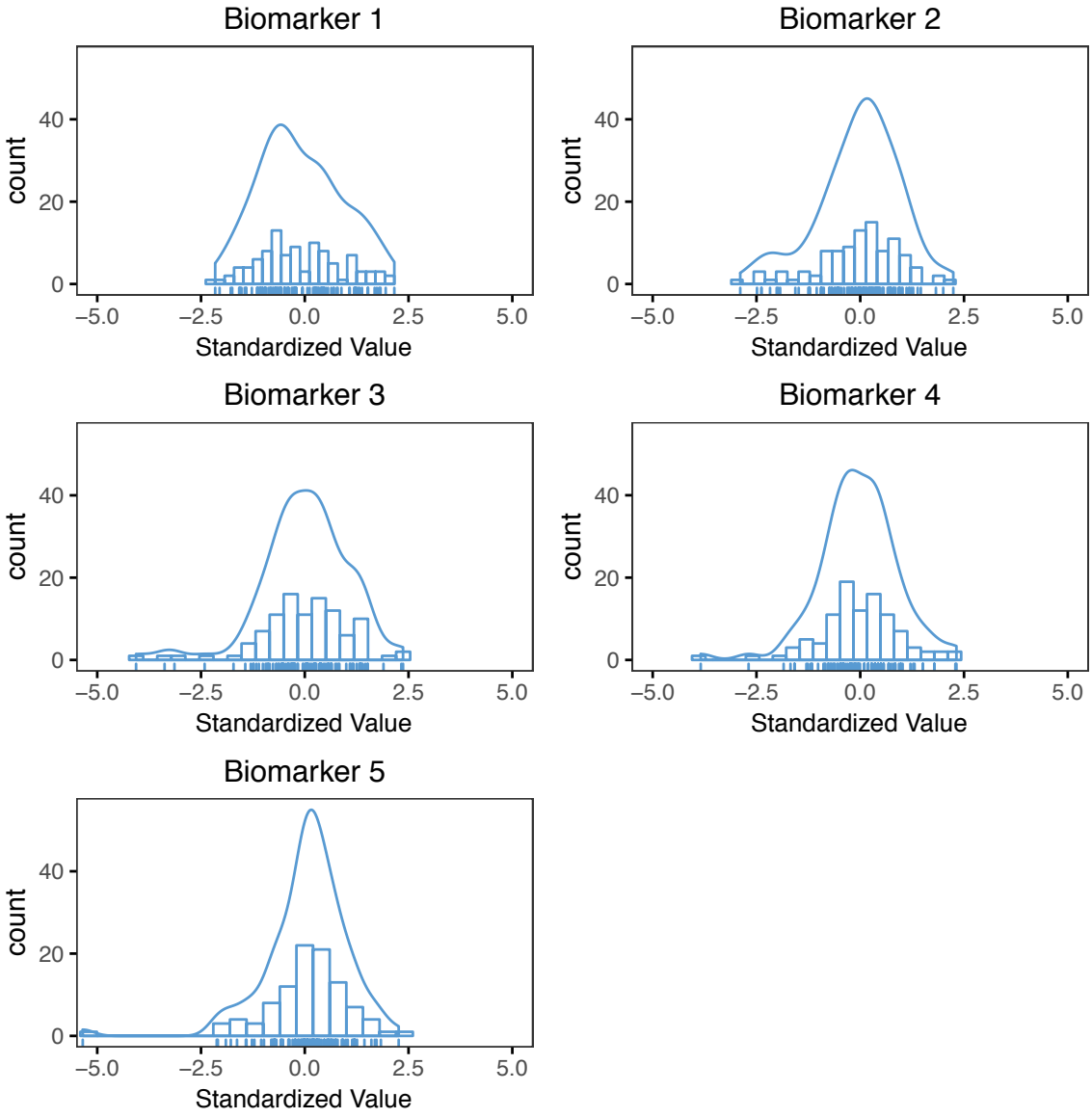

Figure 5: **Cardiovascular Disease Omics Study.** Marginal distributions of the standardized values of each of the five true biomarkers in the Cox proportional hazards model.

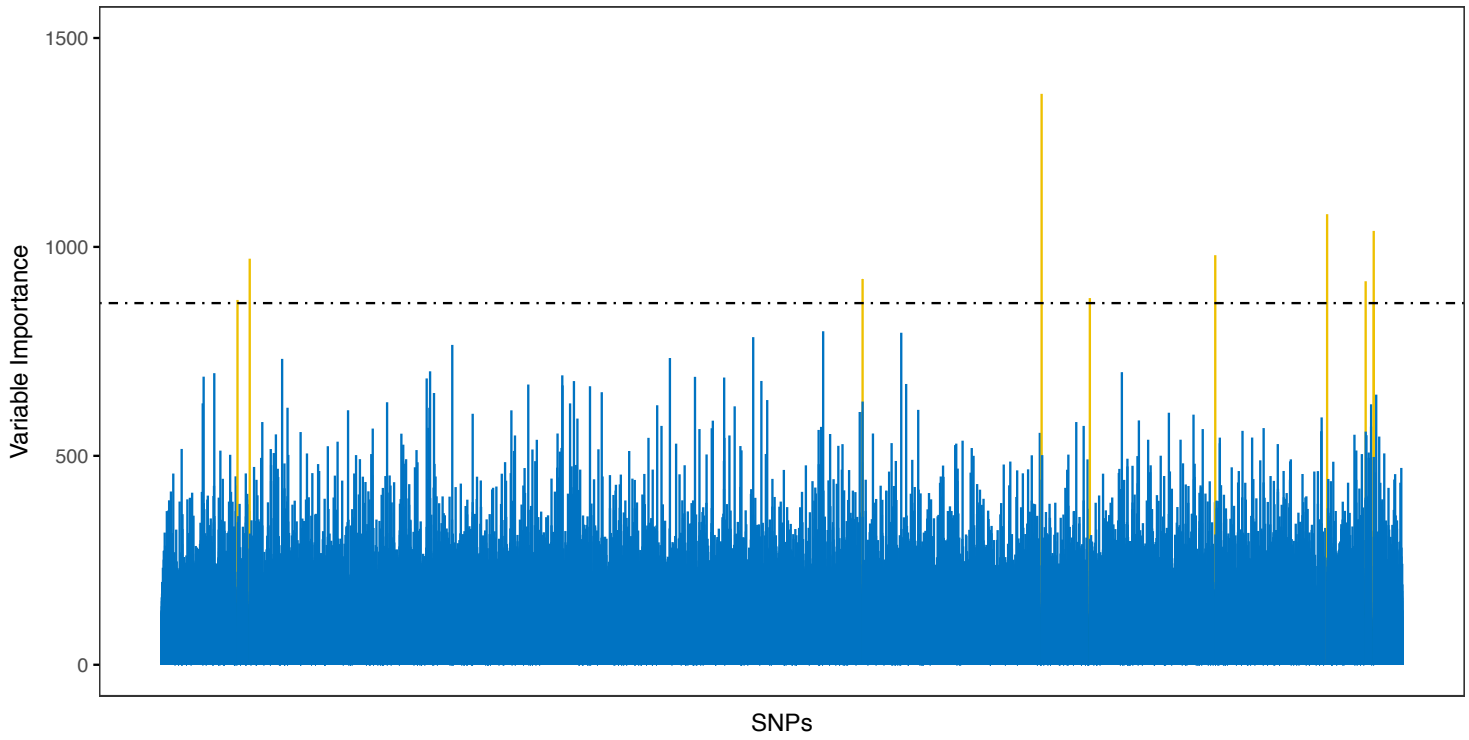

Figure 6: **Women's Health Initiative Genome-wide association study (GWAS) of incident, Type II diabetes.** Bar plot of variable importance for each of 88,277 SNPs. The horizontal dashed line indicates the threshold of selecting top 10 SNPs.