Genomic and Clinical Predictors for Improving Estimator Precision in Randomized Trials of Breast Cancer Treatments: Supplementary Material

Prasad Patil[1], Elizabeth Colantuoni[1], Jeffrey T. Leek[1,*], Michael Rosenblum[1,*]

[1] *Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, U.S.A.*

December 7, 2015

*To whom correspondence may be addressed. 615 N. Wolfe St., Baltimore, MD 21205, mrosen@jhu.edu, jtleek@gmail.com

1 Data Sets GSE19615, GSE11121, GSE7390

The three datasets GSE19615, GSE11121, GSE7390 are available from the Gene Expression Omnibus (Edgar et al., 2002). We obtained the datasets using the MetaGX package in R (available at https://github.com/bhaibeka/MetaGx). Their key characteristics are summarized in Tables 1–3 below. In our analyses, we dropped the two patients in GSE7390 whose tumor grade was unknown.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

2 MammaPrint Prediction

We used the genefu package in R (Haibe-Kains et al., 2012) to make MammaPrint predictions using the gene expression data supplied with each dataset described in Section 1. We specifically used the `gene70` function, which takes as input the expression data matrix and gene annotations and provides as output both a continuous risk score and the dichotomized risk classification. We used the latter as the MammaPrint risk covariate in our covariate adjustment steps. For each dataset, we used the same covariate sets $W_{-ER}, W_C, W_G, W_{CG}$ for adjustment, as defined in section 2.3 of the main text.

3 Differences between unadjusted and adjusted estimators

To assess how different the estimators computed under the unadjusted and adjusted cases are, we looked at the difference $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$ over the $j = 1, \ldots, 100,000$ iterations in each of the four simulations using the four datasets in our study. A histogram of the differences for the simulation using the MammaPrint validation dataset is presented in the main manuscript. Three histograms for the simulations using GSE19615, GSE11121, and GSE7390 appear in this supplement, below. We also present in Table 4 a comparison across all four studies of the average difference, the standard deviation of the difference, and the percentage of times that the unadjusted estimator was larger in absolute value than the adjusted estimator. Since the true treatment effect was set to zero in each simulation study, if the adjustment covariates are prognostic of the outcome and there are substantial chance imbalances between the treatment arms, we would expect the adjusted estimator to be closer to zero more often than the unadjusted estimator. This occurred in over 50% of the iterations in all four studies. In all cases, we used the estimators adjusted for all available covariates (clinical + genomic).

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 4 about here.]

4 Variation in magnitude of precision loss when covariates are not prognostic

We presented in Table 3 of the main text the loss in precision due to adjustment when data was generated from a distribution with W and Y independent. We used more covariates than are usually recommended for this procedure because we wanted to include all clinically relevant baseline covariates that are usually measured for a breast cancer patient. We found that the sample size in the simulated trials and the number of covariates we included affected the magnitude of precision losses. To illustrate, we conducted additional simulation studies where W and Y are independent, both using the MammaPrint validation dataset, where we used fewer adjustment covariates as shown in Table 5. Specifically, we defined new covariate sets W'_{-ER} ={Tumor Size}, W'_C = {Tumor Size, ER status}, W'_G = {MammaPrint Risk Prediction}, W'_{CG} = {Tumor Size, ER status, MammaPrint Risk Prediction}. The precision losses were smaller in magnitude when we reduced the number of adjustment covariates in this way.

[Table 5 about here.]

References

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

Haibe-Kains, B., Schroeder, M., Bontempi, G., Sotiriou, C., and Quackenbush, J. (2012).

genefu: relevant functions for gene expression analysis, especially in breast cancer. r package version 191.

List of Figures

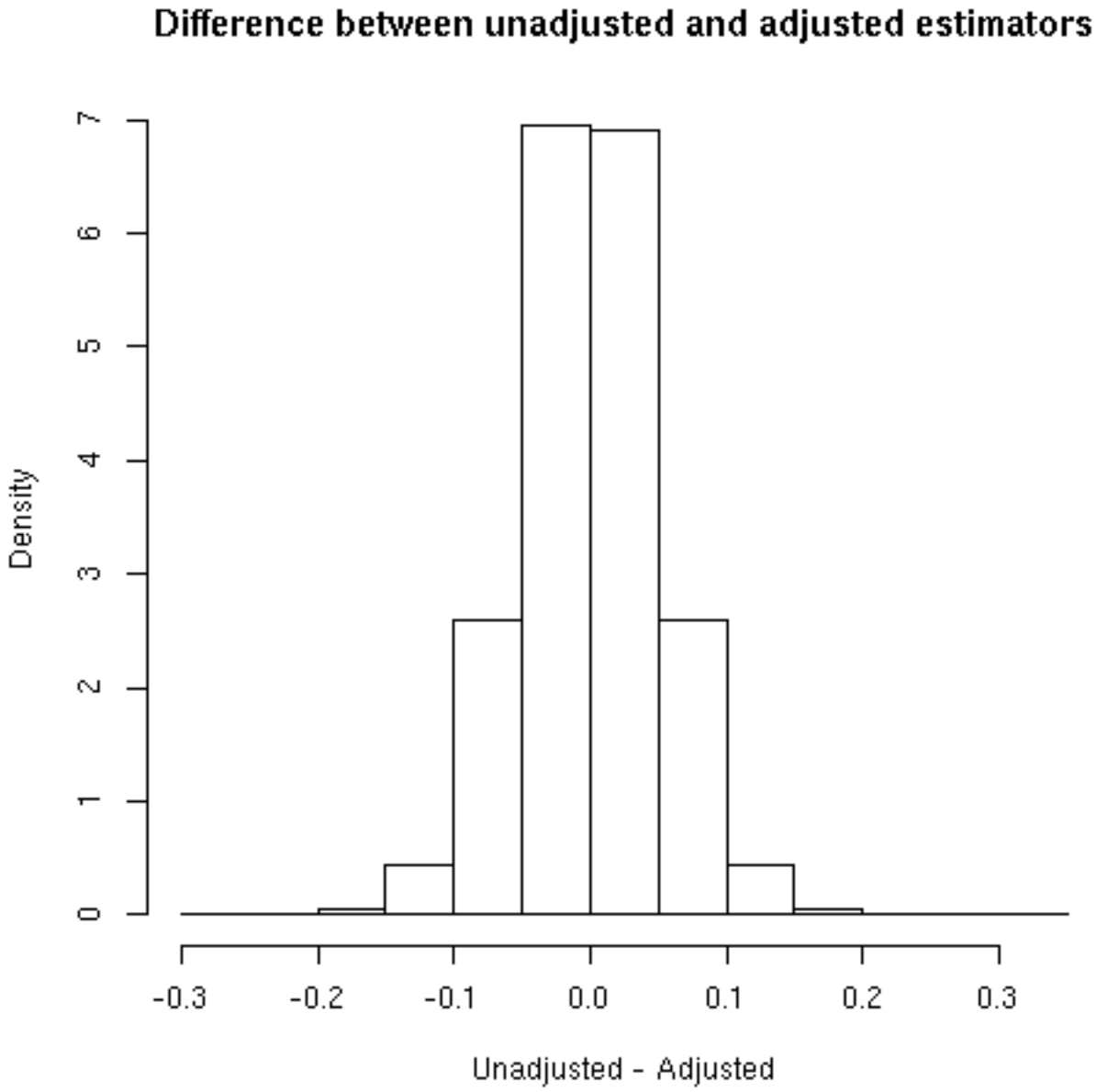

Figure 1: **Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE19615.** The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=-6.7e-07, standard deviation=0.05). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 55% of simulations.

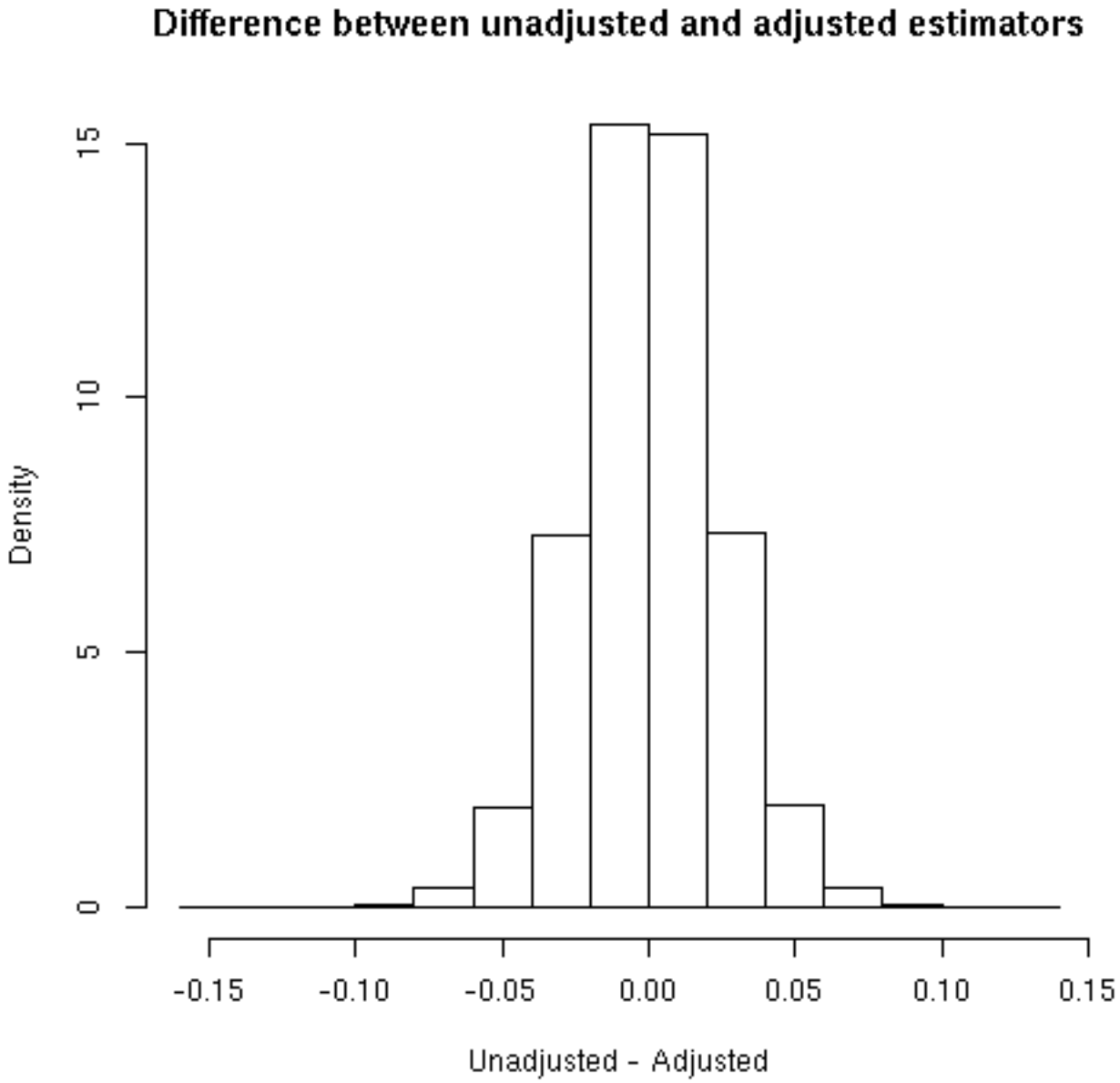

Figure 2: **Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE11121.** The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=-4.6e-05, standard deviation=0.0242). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 53% of simulations.

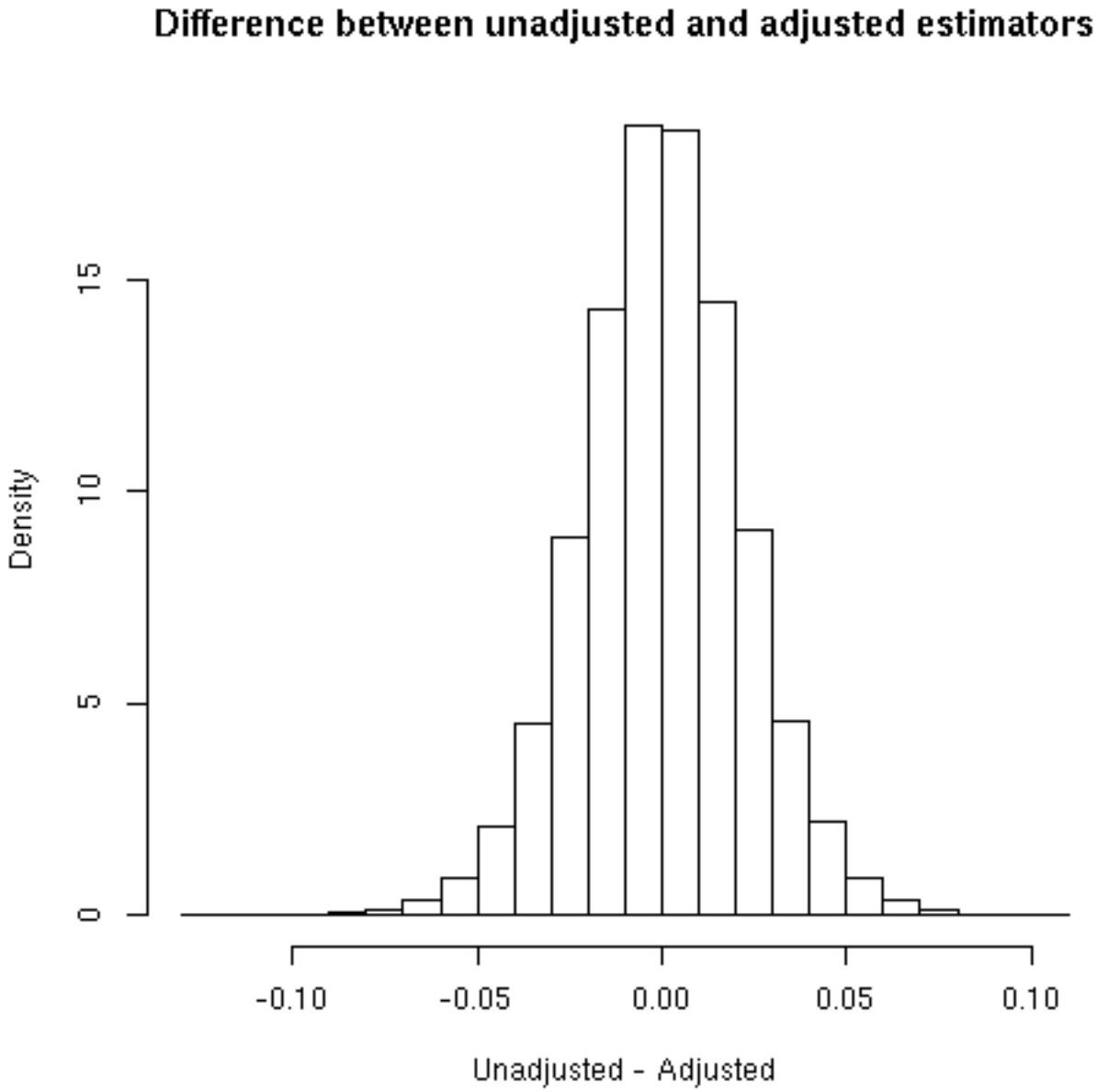

Figure 3: **Histogram of $\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$, GSE7390.** The histogram of differences between the unadjusted and adjusted estimators is roughly normal and is centered close to zero (mean=0.0001, standard deviation=0.0219). The unadjusted estimator is larger in absolute value than the adjusted estimator in approximately 51% of simulations.

List of Tables

Characteristic	Summary
n	115
Age (years)	53.89 (11.78)
Five-Year Recurrence	
Yes	60
No	55
Tumor Size (cm)	2.31 (1.21)
Grade	
1	23
2	28
3	64
Unknown	0
ER	
+	70
−	45
Unknown	0
MammaPrint Risk Prediction	
High	87
Low	28

Table 1: **Characteristics of dataset GSE19615.** ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Characteristic	Summary
n	200
Age (years)	59.98 (12.36)
Five-Year Recurrence	
Yes	153
No	47
Tumor Size (cm)	2.07 (0.99)
Grade	
1	29
2	136
3	35
Unknown	0
ER	
+	162
−	38
Unknown	0
MammaPrint Risk Prediction	
High	142
Low	58

Table 2: **Characteristics of dataset GSE11121.** ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Characteristic	Summary
n	198
Age (years)	46.39 (7.22)
Five-Year Recurrence	
Yes	135
No	63
Tumor Size (cm)	2.18 (0.80)
Grade	
1	30
2	83
3	83
Unknown	2
ER	
+	134
−	64
Unknown	0
MammaPrint Risk Prediction	
High	144
Low	54

Table 3: **Characteristics of dataset GSE7390.** ER - estrogen receptor status, Grade - tumor severity grading (3 is most severe), Five-Year Recurrence - whether or not cancer has reappeared after five years, MammaPrint risk prediction - high or low risk for cancer recurrence. Age and Tumor Size are given as means with standard deviations in parentheses.

Dataset	mean($\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$)	SD($\hat{\psi}^j_{una} - \hat{\psi}^j_{adj}$)	% $\lvert\hat{\psi}^j_{una}\rvert > \lvert\hat{\psi}^j_{adj}\rvert$
Mammaprint	5.4e-05	0.0145	53.1
GSE19615	-6.7e-07	0.05	55.2
GSE11121	-4.6e-05	0.0242	52.8
GSE7390	1.3e-04	0.0219	51.0

Table 4: **Differences between unadjusted and adjusted estimators** We find that the average difference between the unadjusted and adjusted estimators is similar across all simulations and the standard deviations are comparable, although the standard deviation in GSE19615 is more than twice as large as the others. The final column in the table shows the percentage of simulation iterations in which the adjusted estimator was closer in absolute value than the unadjusted estimator to the true treatment effect of zero. For each dataset, this occurred in slightly more than 50% of the iterations.

	Original Sample Size		
Covariate Set	σ^2_{una}	σ^2_{adj}	G_{adj}
W_{-ER}	0.00178	0.00180	-1.1%
W_C	0.00178	0.00181	-1.5%
W_G	0.00178	0.00179	-0.4%
W_{CG}	0.00178	0.00181	-1.8%

Table 5: **Precision gains under data generating distribution with W and Y independent, based on marginal distributions from Mammaprint validation data set, using fewer clinical covariates.**