Measuring the Contribution of Genomic Predictors to Improving Estimator Precision in Randomized trials - Supplementary Material

Prasad Patil, Michael Rosenblum, and Jeffrey T. Leek
Department of Biostatistics, Johns Hopkins University, Baltimore, MD

April 15, 2015

In this supplement, we present an analysis of treatment effect estimator precision gain due to genomic covariates for three breast cancer datasets additional to the one presented in the main text. These additional datasets serve as confirmation that the gains we saw in the MammaPrint data are typical and indicative of the value of the MammaPrint prediction.

1 Data

The three datasets are available from the Gene Expression Omnibus [1]. We obtained the datasets using the MetaGX package in R (available at https://github.com/bhaibeka/MetaGx). The IDs for these datasets are GSE19615, GSE11121, and GSE7390. Their key characteristics are described in **Tables 1,2,3**.

For the analysis, we dropped the two patients in GSE7390 whose tumor grade was unknown.

2 MammaPrint Prediction

We used the genefu package in R [2] 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 outputs both a continuous risk score and the dichotomized risk classification. We used the latter as the MammaPrint risk covariate in our covariate adjustment steps. We used the same covariate sets $W_{-ER}, W_C, W_G, W_{CG}$ for adjustment, as described in section 2.3 of the main text.

3 Results

We applied the covariate adjustment and simulation methods described in sections 2.2 and 2.4 of the main text. We present the results for each dataset as we did in the **Table 2** of the results section.

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: **Baseline characteristics of curated dataset GSE19615** Abbreviations: 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, Tumor Size are given as means with standard deviations.

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: **Baseline characteristics of curated dataset GSE11121** Abbreviations: 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, Tumor Size are given as means with standard deviations.

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: **Baseline characteristics of curated dataset GSE7390** Abbreviations: 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, Tumor Size are given as means with standard deviations.

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
W_{-ER}	0.00247	0.01811	0.00256	0.01678	0.00230	0.01678	7.35%	7.34%
W_C	0.00247	0.01811	0.00269	0.01592	0.00282	0.01579	12.11%	12.82%
W_G	0.00247	0.01811	0.00235	0.01744	0.00234	0.01744	3.71%	3.73%
W_{CG}	0.00247	0.01811	0.00367	0.01640	0.00297	0.01599	9.47%	11.73%

Table 4: **Precision gain under different covariate adjustments - GSE19615** This table presents the simulated estimates for the treatment effect and variance of the treatment effect estimator when unadjusted ($\hat{\psi}_{una}$) and under the two adjustment approaches $\hat{\psi}_{rot}, \hat{\psi}_{col}$. Each of 10,000 times, we resampled records from the original dataset with replacement to generate a new dataset of size $n = 115$. In every iteration, we adjusted the treatment effect estimator using a prespecified set of baseline covariates: W_{-ER} is clinical covariates only, excluding ER status; W_C is all clinical covariates only; W_G is only genomic covariates; W_{CG} includes all clinical and genomic covariates.

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
W_{-ER}	-0.00116	0.00721	-0.00148	0.00659	-0.00131	0.00669	8.57%	7.16%
W_C	-0.00116	0.00721	-0.00145	0.00660	-0.00143	0.00678	8.52%	5.91%
W_G	-0.00116	0.00721	-0.00114	0.00710	-0.00114	0.00710	1.55%	1.55%
W_{CG}	-0.00116	0.00721	-0.00105	0.00651	-0.00130	0.00672	9.76%	6.72%

Table 5: **Precision gain under different covariate adjustments - GSE11121** This table presents the simulated estimates for the treatment effect and variance of the treatment effect estimator when unadjusted ($\hat{\psi}_{una}$) and under the two adjustment approaches $\hat{\psi}_{rot}, \hat{\psi}_{col}$. Each of 10,000 times, we resampled records from the original dataset with replacement to generate a new dataset of size $n = 200$. In every iteration, we adjusted the treatment effect estimator using a prespecified set of baseline covariates: W_{-ER} is clinical covariates only, excluding ER status; W_C is all clinical covariates only; W_G is only genomic covariates; W_{CG} includes all clinical and genomic covariates.

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
W_{-ER}	-0.00091	0.00878	-0.00061	0.00895	-0.00072	0.00895	-1.97%	-1.90%
W_C	-0.00091	0.00878	-0.00086	0.00898	-0.00097	0.00898	-2.28%	-2.27%
W_G	-0.00091	0.00878	-0.00085	0.00844	-0.00085	0.00844	3.86%	3.86%
W_{CG}	-0.00091	0.00878	-0.00127	0.00868	-0.00148	0.00868	1.14%	1.15%

Table 6: **Precision gain under different covariate adjustments - GSE7390** This table presents the simulated estimates for the treatment effect and variance of the treatment effect estimator when unadjusted ($\hat{\psi}_{una}$) and under the two adjustment approaches $\hat{\psi}_{rot}, \hat{\psi}_{col}$. Each of 10,000 times, we resampled records from the original dataset with replacement to generate a new dataset of size $n = 198$. In every iteration, we adjusted the treatment effect estimator using a prespecified set of baseline covariates: W_{-ER} is clinical covariates only, excluding ER status; W_C is all clinical covariates only; W_G is only genomic covariates; W_{CG} includes all clinical and genomic covariates.

We find that gains due to addition of the MammaPrint risk prediction varied from slight loss to 2% gain over using only clinical factors. These results remain in line with what we saw in the main result of the paper.

References

[1] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[2] B Haibe-Kains, M Schroeder, G Bontempi, C Sotiriou, and J Quackenbush. genefu: relevant functions for gene expression analysis, especially in breast cancer. r package version 191, 2012.