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

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

April 21, 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.

We also present results that elucidate how maximum losses in precision when covariates and the outcome are uncorrelated can be affected by how many covariates we use and the size of our dataset.

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.

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.

3 Variation in magnitude of precision loss when covariates are uninformative

We presented in **Table 3** of the main text the loss in precision we observed when we permuted the data and computed the adjusted estimators. This simulation represented what would occur if all covariates included in the adjustment were uncorrelated with the outcome. We saw losses in the 2.5-5% range across all of the sets of covariates. We stress that 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 size of our resampled datasets and the number of covariates we included affected the size of these precision losses in the fully permuted case. To illustrate, we simulated two additional permutation cases, both using the MammaPrint validation dataset: one where we doubled the size of the resampled data (**Table 4**), and one where we used fewer adjustment covariates (**Table 5**). In the fewer covariates case, 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}.

We found that the losses were half as severe when we doubled the sample size or reduced the number of adjustment covariates to two or three.

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
W_{-ER}	-0.00022	0.00087	-0.00023	0.00089	-0.00019	0.00088	-0.02305	-0.01460
W_C	-0.00022	0.00087	-0.00024	0.00089	-0.00022	0.00088	-0.02606	-0.01436
W_G	-0.00022	0.00087	-0.00021	0.00087	-0.00021	0.00087	-0.00266	-0.00266
W_{CG}	-0.00022	0.00087	-0.00032	0.00089	-0.00026	0.00088	-0.02767	-0.01662

Table 4: **Precision gain under different covariate adjustments - full permutation, doubled sample size** 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, all covariate values in the MammaPrint validation data were permuted to simulate independence and a new treatment indicator was randomly generated. This was done to demonstrate that covariates independent of the outcome will provide no precision gain yet will not materially hurt precision. In this case, we resampled with replacement a dataset twice the size of the original. 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.

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.

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
1	-0.00028	0.00177	-0.00035	0.00181	-0.00029	0.00180	-0.01988	-0.01346
2	-0.00028	0.00177	-0.00031	0.00182	-0.00029	0.00181	-0.02770	-0.01852
3	-0.00028	0.00177	-0.00030	0.00178	-0.00030	0.00178	-0.00247	-0.00247
4	-0.00028	0.00177	-0.00029	0.00183	-0.00032	0.00180	-0.03361	-0.01800

Table 5: **Precision gain under different covariate adjustments - full permutation, fewer covariates** 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, all covariate values in the MammaPrint validation data were permuted to simulate independence and a new treatment indicator was randomly generated. This was done to demonstrate that covariates independent of the outcome will provide no precision gain yet will not materially hurt precision. 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.

[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.