Measuring the Contribution of Genomic Predictors for 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

June 22, 2015

We present an analysis of treatment effect estimator precision gain due to genomic covariates for three additional breast cancer datasets. These additional datasets serve as confirmation that the gains we saw in the MammaPrint data may be typical and indicative of the value of the MammaPrint prediction.

We also show how much variability in the approximate gains we might expect if the dataset were slightly different than the original. This is done by taking a 75% resampling of the data for recalculating the precision gain.

Finally, we present results that elucidate how 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 (Edgar et al., 2002). 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, and 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 (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 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. 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.

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 The impact of differences in the original data

To examine the impact of having a slightly different dataset than the original dataset, we created 10 modified datasets, each consisting of a random subset of 75% of the original participants from the MammaPrint validation data (where each participant's data vector was kept intact). For each modified dataset we re-ran the entire analysis: resampling 10,000 simulated trials each with 296 participants, as described at the end of section 2.4. Histograms of the gains (G_{col}, the precision gain from the adjusted estimator $\hat{\psi}_{col}$) due to clinical factors and the gains due to adding the genomic predictor appear in **Figure 1**. We found that the approximation of the gain may vary ±3% around its mean, suggesting that our approximations from the resampling of the original datasets are fairly stable.

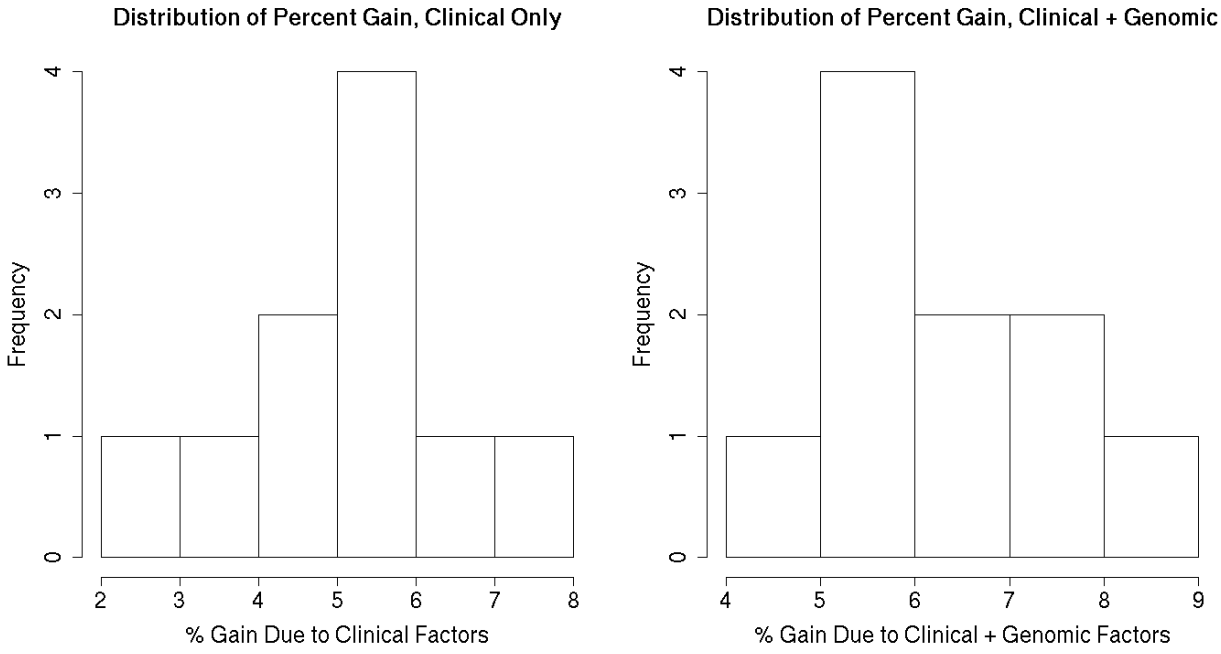

Figure 1: **Variability in Precision Gain in MammaPrint Validation Data** For the simulation studies based on the ten subsamples, the left panel shows the histogram of the gain (G_{col}) due to clinical covariates only (W_C), and the right panel shows the gain due to both clinical and genomic covariates together (W_{CG}). The gains vary ±3% from the original approximations.

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 saw precision losses from 0.3% to 5.4% across the sets

of covariates. 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 two additional simulation studies where W and Y are independent, both using the MammaPrint validation dataset: one where we doubled the simulated trial sample size (**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 roughly 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}	-0.0001	0.00090	-0.0001	0.00091	-0.0001	0.00091	-1.8%	-1.1%
W_C	-0.0001	0.00090	-0.0001	0.00092	-0.0001	0.00091	-2.2%	-1.2%
W_G	-0.0001	0.00090	-0.0001	0.00090	-0.0001	0.00090	-0.2%	-0.2%
W_{CG}	-0.0001	0.00090	-0.0001	0.00092	-0.0001	0.00091	-2.6%	-1.5%

Table 4: **Precision gains under data generating distribution with W and Y independent, doubled sample size**

Covariates	B_{una}	σ^2_{una}	B_{rot}	σ^2_{rot}	B_{col}	σ^2_{col}	G_{rot}	G_{col}
W_{-ER}	0.0001	0.00178	0.0001	0.00181	0.0001	0.00180	-1.8%	-1.1%
W_C	0.0001	0.00178	0.0001	0.00182	0.0001	0.00181	-2.5%	-1.5%
W_G	0.0001	0.00178	0.0001	0.00179	0.0001	0.00179	-0.4%	-0.4%
W_{CG}	0.0001	0.00178	0.0001	0.00183	0.0001	0.00181	-3.0%	-1.8%

Table 5: **Precision gains under data generating distribution with W and Y independent, fewer covariates**

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.