A Statistical Reappraisal of the Findings of an Esophageal Cancer Genome-Wide Association Study

To the Editor:

A recent esophageal cancer genome-wide association study by Hu and colleagues (1) identified 37 statistically significant single nucleotide polymorphisms (SNP) and reported a nearly perfect classification of cancer cases and controls on the basis of only these SNPs. Taken at face value, this implies that esophageal cancer is a solely genetic disease, although literature in the field suggests that environmental factors make a major contribution to susceptibility for many cancer types (2). To shed light on this issue, we reanalyzed the data of Hu and colleagues (1) and identified two data analysis pitfalls that caused overoptimistic conclusions in the original article.

First, the SNP selection method by Hu and colleagues (1) was severely biased toward claiming significance for SNPs that are not truly associated with the disease. The calculation of *P* value in the published generalized linear model (GLM)–based SNP selection method does not reflect the significance of the SNP under consideration but the significance of three variables combined (SNP, family history of esophageal cancer, and alcohol consumption). Because family history and alcohol consumption are strong risk factors for esophageal cancer, this *P* value will be biased toward zero, even when the SNP has nothing to do with esophageal cancer. When an unbiased GLM-based procedure is used instead, no SNPs can be found significant at the Bonferroni adjusted 0.05 α-level. See Fig. 1 for details and histograms of the distributions of SNP *P* values produced by both previously published and unbiased procedures for SNP screening.

Second, both SNP selection and building of the principal component analysis–based classifier model were performed by Hu and colleagues (1) on the same 100 subjects as used for estimation of the final classification accuracy. Because neither cross-validation nor independent sample validation was performed, the resulting classification performance estimate is overoptimistic as explained by Simon and colleagues (3). To obtain an unbiased performance estimate for the SNP selection method and the classifier of Hu and colleagues (1), the above methods were applied by repeated 10-fold cross-validation procedure (4). The resulting classification performance estimate was 0.68 area under receiver-operating characteristics curve (AUC), whereas the original procedure in Hu and colleagues (1) led to 0.98 AUC, indicating a 0.30 AUC overestimation.

These findings suggest that the data analysis of Hu and colleagues (1) identified nonstatistically significant SNPs and derived a severely biased estimate of classification performance of esophageal cancer patients and healthy controls. For a study of effects of environment and genetics versus data analysis pitfalls, see Statnikov and colleagues (5). The present case study also underscores the importance of sound data analysis in genome-wide association studies.

Alexander Statnikov
Discovery Systems Laboratory,
Department of Biomedical Informatics,
Vanderbilt University,
Nashville, Tennessee

Chun Li
Department of Biostatistics,
Center for Human Genetics Research,
Vanderbilt University,
Nashville, Tennessee

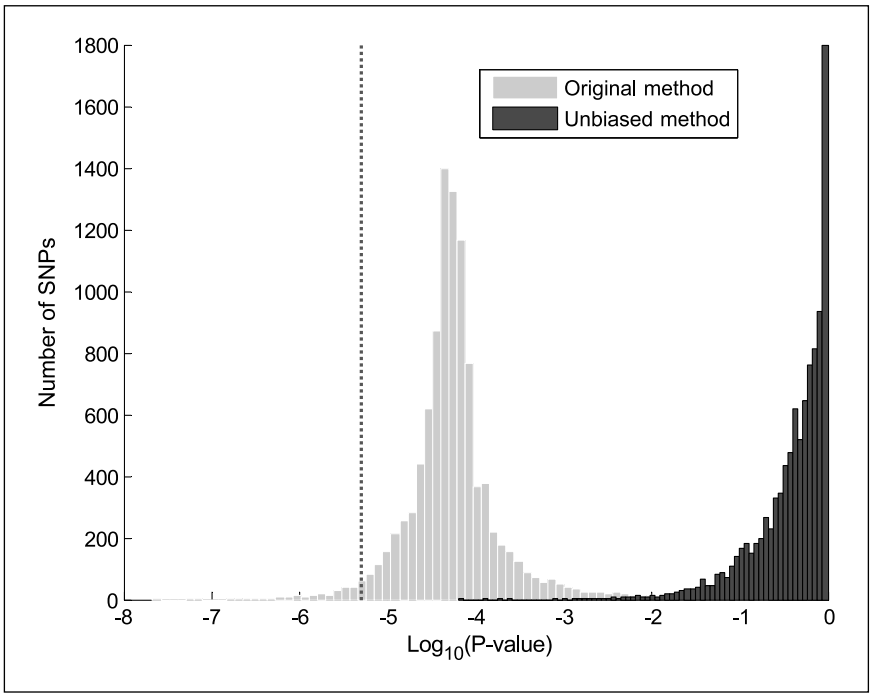

Figure 1. Distributions of SNP *P* values computed by the original GLM-based method of Hu and colleagues (1) and the unbiased procedure. *Dotted vertical line,* Bonferroni adjusted α-level (0.05/10k). The original method used the model with no independent variables as the "null" versus that including alcohol consumption, family history, and SNP. In contrast, the unbiased method tests the addition of SNP to a model with alcohol consumption and family history. Although there are SNPs that are significant according to the original method, no SNP is significant by the unbiased method.

Constantin F. Aliferis
Discovery Systems Laboratory,
Departments of Biomedical Informatics, Biostatics, and Cancer Biology,
Vanderbilt University,
Nashville, Tennessee

References

1. Hu N, Wang C, Hu Y, et al. Genome-wide association study in esophageal cancer using GeneChip mapping 10K array. Cancer Res 2005;65:2542–6.
2. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer 2002;99:260–6.
3. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003; 95:14–8.
4. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004;20:374–80.
5. Statnikov A, Li C, Aliferis CF. Effects of environment, genetics and data analysis pitfalls in an esophageal cancer genome-wide association study. PLoS ONE 2007;2:e958.

doi:10.1158/0008-5472.CAN-07-2999