Re: Lessons from Controversy: Ovarian Cancer Screening and Serum Proteomics

We read with interest the paper by Ransohoff *(1)*, which addresses the important issue of validity of published results in "-omics" fields to assess molecular markers for diagnosis and prognosis of cancer. The commentary was directly motivated by the debate around an initial study that used mass spectroscopy data to discriminate between samples from patients with ovarian cancer and samples from healthy individuals, with reported sensitivity and specificity of nearly 100%. It discussed many relevant aspects of "-omics" research, including reproducibility and the nature of observational studies. Although there are various issues specific to the new technologies, we would like to add that other issues, including the need for external data in validation, are reasonably well understood in the more traditional context of prognostic models *(2)*.

The debate itself concerns prediction error rate directly. The true error rate is unknown and has to be estimated from the data at hand. The performance of estimators of the unknown prediction error rate, however, seems to have been discussed only relatively recently in the "-omics" literature *(3)*. The key issues surrounding the particular debate described by Ransohoff *(1)* are perhaps not the underestimation of the prediction error rate from a narrow statistical point of view, but we argue additionally that some indication of the credibility of the reported prediction error rate, or reported findings in general, should be given in a publication, whenever possible. This is necessary because the amount of intrinsic information in data is critical in determining the validity of a model built from the data *(2)*.

In most biomedical publications, variability measures, such as the standard error or confidence interval, almost always accompany published results. However, "-omics" publications have rarely included variability measurements. Xu and Li *(4)* used the bootstrap method to estimate the "rediscovery rate" of the top ranking genes that are differentially expressed under different experimental conditions. Given a target number of genes χ, the rediscovery rate is defined as the percentage of the original top χ genes reported from the data that will be rediscovered as top χ genes if the experiment is to be independently replicated. The rediscovery rate can be seen as a measure of variability for the gene ranking problem. In their example of 28 Affymetrix chips to distinguish between two different treatments, the estimated rediscovery rates are about 10% and 53%, respectively, for $\chi \leq 100$, by use of two different methods to select the top ranking genes. The bootstrap method used to estimate the rediscovery rate in this case is justified as described by Efron and Mammen *(5,6)*. Bootstrap can also be used to give estimates of variability (e.g., confidence intervals) for prediction error rate *(7)*. We recommend that the associated variability be always reported as part of the result.

RONGHUI XU
ANTHONY GAMST

REFERENCES

(1) Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97:315–9.
(2) Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.
(3) Braga-Neto UM, Dougherty ER. Is cross validation valid for small sample microarray classification? Bioinformatics 2004;20:374–80.
(4) Xu R, Li X. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. Bioinformatics 2003;19:1284–9.
(5) Efron B, Tibshirani R. The problem of regions. Ann Stat 1998;26:1687–718.
(6) Mammen E. Bootstrap and wild-bootstrap for high-dimensional linear models. Ann Stat 1993;21:255–85.
(7) Efron B, Tibshirani R. Improvements on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.

NOTES

Affiliation of authors: Rebecca and John Moores Cancer Center, Department of Family and Preventive Medicine and Department of Mathematics, University of California, San Diego.

Correspondence to: Ronghui Xu, PhD, Moores Cancer Center, 9500 Gilman Dr., Mail Code 0112, La Jolla, CA 92093-0112 (e-mail: rxu@ucsd.edu).

DOI: 10.1093/jnci/dji235

RESPONSE

Xu and Gamet note that confidence intervals, often reported in research results, are seldom used in studies of diagnosis, though they have recently been advocated *(1)*. Reporting confidence intervals could indicate an important kind of numerical uncertainty around the "central tendency" of a result, but it would not, as Xu correctly notes, address larger problems of bias and overfitting, and it would not obviate the need to demonstrate reproducibility in an independent group *(2,3)*.

DAVID F. RANSOHOFF

REFERENCES

(1) Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1–12.

(2) Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4:309–14.

(3) Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142–9.

NOTES

Correspondence to: David F. Ransohoff, MD, Departments of Medicine and Epidemiology, University of North Carolina at Chapel Hill, CB #7080, Bioinformatics Bldg. 4103, Chapel Hill, NC 27599-7080 (e-mail: ransohof@med.unc.edu).

DOI: 10.1093/jnci/dji236