BRIEF COMMUNICATION

Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer

Keith A. Baggerly, Jeffrey S. Morris, Sarah R. Edmonson, Kevin R. Coombes

Proteomic profiling of serum initially appeared to be dramatically effective for diagnosis of early-stage ovarian cancer, but these results have proven difficult to reproduce. A recent publication reported good classification in one dataset using results from training on a much earlier dataset, but the authors have since reported that they did not perform the analysis as described. We examined the reproducibility of the proteomic patterns across datasets in more detail. Our analysis reveals that the pattern that enabled successful classification is biologically implausible and that the method, properly applied, does not classify the data accurately. We show that the method used in previously published studies does not establish reproducibility and performs no better than chance for classifying the second dataset, in part because the second dataset is easy to classify correctly. We conclude that the reproducibility of the proteomic profiling approach has yet to be established. [J Natl Cancer Inst 2005;97:307–9]

Dramatic results *(1)* from proteomic profiling of serum by use of mass spectrometry have triggered hopes that this technology will provide a diagnostic test for ovarian cancer *(2,3)*. In this approach, proteomic patterns—i.e., the joint intensities of several spectral peaks—are used to distinguish samples obtained from patients with ovarian cancer or from healthy individuals *(1)*.

Diagnostic application of this approach requires that patterns from previous studies suffice to classify new spectra. Proteomic patterns in general, however, have been difficult to reproduce *(4)*. Zhu et al. *(5)* reported that, by use of a new method, a pattern derived from one ovarian cancer dataset accurately classified a second, blinded dataset produced many months later. These results are in contrast with suggestions that a systematic measurement offset between these particular datasets precludes reproducibility *(6)*.

However, a programming error precluded the classification across datasets as reported. According to the reported method *(5),* the first dataset was split into training and test sets. A separating pattern of 18 peaks (at *m*/*z* values 167.8031, 321.42, 322.42, 359.63, 385.57, 413.17, 433.91, 434.69, 444.47, 445.26, 1222.18, 1528.34, 3345.80, 3349.15, 3473.31, 3528.53, 6101.63, and 6123.52) was identified in the training set and confirmed in the test set. Spectra from the second dataset were classified according to the identity of the five nearest neighbors [by Mahalanobis distance *(7)*] among the training spectra set from the first dataset. In reality, however, "the spectra in the second dataset were classified using a jack-knife approach where distances were computed between each spectrum and all of the other spectra in the second dataset, and the spectrum was classified according to the status of its five nearest neighbors in this set of spectra. Only the peak locations (*m*/*z* values) were retained across datasets, and these served to define the points at which the distances were computed. Further, the validation simulations used training sets drawn from the second dataset" (Wei Zhu, personal communication).

This creates a problem for the claim of reproducibility across datasets, because classification of the second dataset used knowledge of the status of spectra in the second dataset. Nonetheless, the separation achieved suggests that these 18 peak *m*/*z* values may be "important" for classification.

In this study, we investigate the biological plausibility of the reported *m*/*z* values for cancer diagnosis, and then we compute classification rates obtained with their reported values or with newly generated patterns. We then replicate the jack-knife approach, as described (Wei Zhu, personal communication), to classify the second dataset by use of the published *m*/*z* values listed above. Finally, we calculate the probability that the reported classification could occur by chance.

Zhu et al. *(5)* analyzed two publicly available datasets *(8)*. The first dataset [Ovarian Cancer Dataset 4–3–02 *(8)*] contains 216 spectra, obtained from the serum of 100 cancer patients and 116 "unaffected" patients; 100 of the latter were obtained from healthy control subjects, and the remaining 16 were from patients with benign ovarian disease. The second dataset [Ovarian Cancer Dataset 8–7–02 *(8)*] contains 253 spectra, obtained from the serum of 162 cancer patients and 91 healthy control subjects. The Matlab code for all analyses is available (http://bioinformatics.mdanderson.org).

To address the biological plausibility of the 18 identified peaks, we computed two-sample *t* statistics by comparing all control samples with all cancer samples at each peak separately for the two datasets. If cancer-induced changes in protein expression are measurable at those *m*/*z* values, then the *t* statistics should have the same sign in both datasets.

To test the published classification method for the second dataset, we drew training sets of 50 cancer sample spectra and 50 unaffected control sample spectra from the first dataset. For each of the 18 reported peaks, we classified each spectrum in the second dataset as cancer or control by use of the 5 nearest neighbors in the training set. We repeated this procedure 1000 times.

Next, we used randomly chosen training sets from the first dataset to define peak sets by the approach described by Zhu et al. *(5)* and then used these patterns to classify the second dataset as above. As before, we repeated this procedure 1000 times.

Affiliations of authors: Department of Biostatistics, M.D. Anderson Cancer Center, Houston, TX (KAB, JSM, KRC); Department of Family and Community Medicine, Baylor College of Medicine, Houston, TX (SRE)

Correspondence to: Keith A. Baggerly, Department of Biostatistics, 1515 Holcombe Blvd., Box 447, Houston, TX 77030–4009 (e-mail: kabagg@mdanderson.org).

See "Notes" following "References."

DOI: 10.1093/jnci/dji008

Journal of the National Cancer Institute, Vol. 97, No. 4,

Finally, after replicating the jack-knife classification of the spectra in the second dataset using the published *m/z* values (derived from the first dataset) listed above, we randomly chose sets of 18 peaks and classified the second dataset spectra by this jack-knife approach. We repeated this procedure 1000 times. We then repeated the jack-knife process, first with random peaks chosen from *m/z* values of less than 6000 and second with *m/z* values of less than 1000, to reflect the range of most of the original 18 peaks found by Zhu et al.

The signs of the *t* statistics changed between datasets for 13 of the 18 peaks (Fig. 1). A change in sign indicates that protein intensities at that point are higher in cancer spectra in one dataset and in control spectra for the other dataset. This reversal is not consistent with a persistent difference in protein expression between cancer samples and control samples.

Results of simulations (Fig. 2) using the nearest neighbor method, as described by Zhu et al. *(5)*, showed lower accuracy than was reported in that publication. In 893 of 1000 simulations using the published pattern derived from the first dataset, all 253 spectra were classified as cancer. This corresponds to a test with 100% sensitivity but 0% specificity. The highest overall accuracy observed (200 of the 253 spectra) was less than 80%. In 667 of 1000 simulations using patterns newly generated from random training sets, all 253 spectra were classified as control, and in another 218 of the 1000 simulations, all 253 spectra were classified as cancer. The highest overall accuracy observed (172 of the 253 spectra) was less than 70%.

In our hands, application of the jack-knife approach to the second dataset, using the published *m/z* values *(5)*, resulted in correct classification of 249 (98.42%) of the 253 spectra. Classification accuracy using the jack-knife approach with randomly chosen patterns was quite high; in fact, random values met or exceeded 98.4% classification accuracy 6% of the time using the whole spectrum, 14.8% of the time with random *m/z* values of less than 6000, and 56.2% of the time with random *m/z* values of less than 1000.

The pattern of protein expression is inconsistent between the datasets at the reported *m/z* values. Thus, these values apparently do not represent biologically important changes in cancer patients.

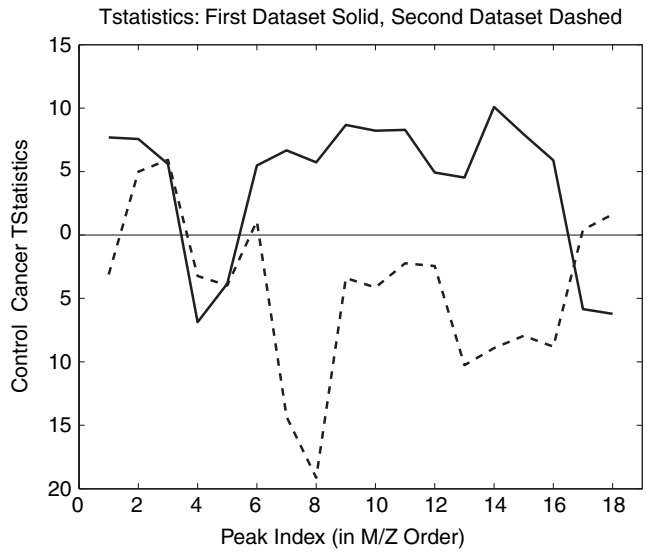

Fig. 1. Summary of *t* statistics at 18 published peaks. Peaks have *m/z* values as indicated in the text. The *t* statistics represent the difference in spectral intensity between cancer and unaffected spectra for the 18 reported *m/z* values. **Solid line** = *t* statistic values from the first dataset; **dashed line** = *t* statistic values from the second dataset. The magnitude and sign of the *t* statistics correspond to the relative protein expression of cancer and normal spectra for the two datasets; a change in sign indicates that the average spectral intensity at that *m/z* value was greater in cancer spectra for one dataset and for control spectra in the other.

Further, neither the method outlined in Zhu et al. *(5)* nor the jack-knife method demonstrate reproducibility in these datasets. The former, theoretically a valid test of reproducibility, results in unacceptably poor classification. The latter does not diagnose new cases on the basis of the previous data only and results in

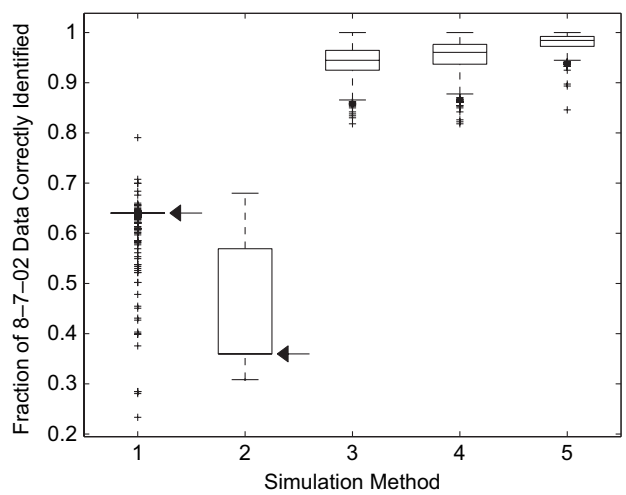

Fig. 2. Classification accuracies observed in simulations. **Box plots** show the median and quartile accuracies observed for each simulation approach. Each simulation involved 1000 repetitions. Simulation methods are as follows: Method 1) Training sets randomly chosen from the first dataset were used to classify the second dataset according to the published method using the 18 *m/z* values listed in the text. The **arrow** indicates the median line, also the first and third quartiles, which coincide with the observed accuracy when all samples are classified as "cancer." Method 2) Training sets randomly chosen from the first dataset were used to generate new sets of *m/z* values, and these values were used to classify the second dataset according to the published method *(5)*. The **arrow** points to the median line, also the first quartile, which coincides with the observed accuracy when all samples are classified as "control." Method 3) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from the entire spectrum. Method 4) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from values of less than 6000. All of the originally reported *m/z* values were less than 6000. Method 5) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from values of less than 1000. Of the originally reported *m/z* values, 10 of the 18 values were less than 1000.

classifications that are no better than chance.

The excellent classification achieved in the second dataset using random patterns suggests pervasive differences between cancer and control spectra. Changes in protein expression associated with cancer should affect only a few specific peaks, not the entire spectrum. Systematic differences in spectra are more likely associated with procedural bias, such as incomplete randomization, that confounds our ability to recognize potentially reproducible biological factors. Hence, reproduction of proteomic patterns across experiments remains an open question that, in our assessment, has not been answered with the two datasets investigated.

References

(1) Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.

(2) Correlogic Systems: Patterns for Life. http://www.correlogic.com. [Last accessed: November 22, 2004.]

(3) Pollack A. A new cancer test stirs hope and concern. New York Times February 3, 2004.

(4) Rogers MA, Clarke P, Noble J, Munro NP, Paul A, Selby PJ, et al. Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. Cancer Res 2003;63:6971–83.

(5) Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS. Detection of cancer-specific markers amid massive mass spectral data. Proc Natl Acad Sci U S A 2003;100:14666–71.

(6) Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments. Bioinformatics 2004;20:777–85.

(7) Mardia K, Kent J, Bibby J. Multivariate analysis. New York (NY): Academic Press, Harcourt Brace Jovanovich; 1979.

(8) Clinical Proteomics Program Databank. http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. [Last accessed: October 26, 2004.]

Notes

The authors thank Wei Zhu for useful discussions. JSM was supported in part by NIH-NCI RO1 CA 107304.

Manuscript received June 3, 2004; revised October 7, 2004; accepted November 1, 2004.