

# Signal in Noise: Can Experimental Bias Explain Some Results of Serum Proteomics Tests for Ovarian Cancer?

Keith A. Baggerly<sup>a\*</sup>, Jeffrey S. Morris<sup>a</sup>, Sarah R. Edmonson<sup>b</sup> and Kevin R. Coombes<sup>a</sup>

<sup>a</sup>Department of Biostatistics and Applied Mathematics, U.T. M.D. Anderson Cancer Center, Houston, TX

<sup>b</sup>Department of Family and Community Medicine, Baylor College of Medicine, Houston, Texas

Running Head: Signal, Noise, and Randomization

\* To whom correspondence should be addressed:

Department of Biostatistics and Applied Mathematics

U.T. M.D. Anderson Cancer Center

1515 Holcombe Blvd, Unit 447

Houston, TX 77030-4009

Phone: (713) 563-4290

Fax: (713) 563-4243

Email: [kabagg@mdanderson.org](mailto:kabagg@mdanderson.org)

Abstract:

Mass spectrometry-based protein profiling of serum has been suggested as a potential screening test for ovarian cancer. While it has proven difficult to generalize the results from one experiment to another, a recent report demonstrated that a pattern of 18 peaks derived from one dataset proved effective for blinded diagnosis in an unrelated dataset. This persistence, which implies the presence of clinically important biological structure, apparently contradicts published suggestions that the results may be due to experimental bias. We reexamined the data to resolve the apparent contrast.

We used two-sample t-statistics to assess whether the peaks used showed similar differences in protein expression for cancer and control samples in both datasets. Next, we simulated using randomly-chosen protein patterns for prediction, testing whether the reported classification accuracy for the blinded dataset was better than would be expected by chance.

T-tests on all 18 peaks showed significant differences between cancer and control in both datasets. However, the direction of the difference changed for 13 of the peaks; expression levels were higher for cancers in one dataset and for controls in the other. Next, our simulations showed that randomly chosen sets of 18 peaks provided a median classification accuracy equal to the reported results, if the peaks were drawn from the lower mass end of the spectrum.

Both analyses suggest that the observed separation has no biological basis. The degree of separation observed with random patterns suggests experimental bias. We discuss potential causes and implications of this bias, and comment on the need for careful experimental design.

## Introduction: Background and Purpose

Ovarian cancer is common and dangerous: Globally, this cancer has the sixth highest incidence and the seventh highest cancer mortality among women, and five-year relative survival rates remain below 50% even in developed nations (1). Mortality rates are dramatically higher for late-stage diagnosis, and malignant ovarian tumors are typically asymptomatic until the disease is quite advanced. Thus, an effective screening blood test for the early detection of ovarian cancer might dramatically reduce the morbidity and mortality associated with this disease.

Recently, mass spectrometry-derived serum protein profiles have been suggested as just such a cancer screening test. In this technique, detailed in figure 1, spectra are generated from serum samples; peaks in the spectra correspond to particular proteins. Sets of spectral intensities at several distinct  $m/z$  values define “proteomic patterns” for the corresponding biological samples. Because these proteomic patterns ideally represent the relative concentrations of different proteins in the sample, they can be used as biomarkers to indicate the presence or absence of a target disease such as ovarian cancer. The underlying assumption - that a solid, organ-bound tumor may uniquely impact the protein profile in the blood serum - seems reasonable, given that protein markers in the blood have been identified in other neoplastic diseases.

In the initial report on this approach, which appeared in *The Lancet* in 2002 (2), spectra were produced from subjects with ovarian cancer, healthy control subjects, and subjects with benign disease. Half of the cancer and control samples were used as a training set to determine a pattern of  $m/z$  values to be used to discriminate cancer from control; this pattern was demonstrated to predict the status of the remaining samples with

impressive accuracy. In a laudable move, the group reporting the results, the NCI/FDA Clinical Proteomics Program, posted the raw data (dataset 1) underpinning their results on their website. The group has also posted two additional SELDI ovarian cancer datasets (dataset 2 and dataset 3), and the results of these experiments continue to support their findings, as summarized in table 1. These latter two datasets are of particular interest because their spectra – drawn from different sets of women but analyzed using the same type of SELDI chip – can theoretically be compared to check that patterns derived from one dataset still work in the other, thus demonstrating reproducibility of the method. Dataset 1, analyzed using a chip no longer being manufactured, has no such comparable dataset. While there have now been other reports using this approach to biomarker discovery for other types of disease (3-8), in large part the raw data from these experiments is not as readily available.

However, both the proteomic profiling approach and the ovarian cancer datasets are the subject of some controversy, as detailed in table 2. There may be fundamental problems with the technique itself (9), and with reproducibility of SELDI data over longer periods of time (10). In addition, the ovarian cancer datasets have demonstrable flaws (11, 12), suggesting that the reported success at separating cancer from control may reflect experimental bias rather than true biological differences between the groups. In particular, Baggerly et al. (11) note flaws in data calibration that make it extremely unlikely that biologic patterns observed in dataset 2 will be reproduced in dataset 3. If that is so, the proteomic patterns used in these datasets will not be useful as a diagnostic tool, because their success at separating cancer from normal spectra cannot be demonstrated reproducibly.

On the other hand, an report by Zhu et al. (13) appears to refute these concerns by demonstrating that a proteomic pattern of 18 peaks chosen to be diagnostic for dataset 2 also proved effective for classifying blinded spectra from dataset 3, with reported sensitivities and specificities near 100% for both datasets. This latter report has been cited as evidence for the presence of persistent and clinically important biological structure in these two datasets (14).

The apparent conflict between Baggerly et al.'s (11) prediction of the lack of consistent structure and Zhu et al.'s (13) finding of consistent structure has attracted considerable interest (14-16). This latter report - that a discrimination pattern can be generalized to at least two datasets – represents the strongest evidence to date that findings from the ovarian datasets may represent biologically relevant, generalizable information. Thus, reproduction and close examination of this work is a reasonable approach to determining whether these findings should be extended to clinical applications.

In this paper, we reexamine the pattern found to give persistent separation between cancers and controls across datasets 2 and 3. This persistent pattern is assessed according to two criteria: biological plausibility and statistical significance.

With respect to biological plausibility, we note that for the proteomic patterns to be clinically useful, these patterns should represent consistent underlying biological processes. Otherwise, we cannot expect that the findings from one patient set or one laboratory to be useful for other labs and other patients – so we cannot depend on the results in a clinical application. At a minimum, we should expect that where spectral peaks show clear differences between cancer and control, the difference between cancer

and control should follow the same direction across datasets. A peak showing stronger expression for cancers than controls in one dataset should also be stronger for cancers in another dataset. In order to judge biological plausibility, we examine two-sample T-statistics contrasting cancers with controls at the 18 peaks in the initial and blinded datasets.

With respect to statistical significance, we note that the classification accuracy of the proteomic pattern applied to the blinded dataset is quite high. The question is whether this level of accuracy is unexpected. At a minimum, we should expect a carefully chosen proteomic pattern to provide “significantly” better classification accuracy than a pattern chosen at random. In order to judge statistical significance, we simulate large numbers of random proteomic patterns, use them to classify cancers and controls in the blinded dataset, and compare the simulated values with the observed classification accuracy. Finally, we discuss the implications of our findings and offer some additional comments on the importance of careful experimental design.

## Materials and Methods

### The Data

Zhu et al.’s work concerned two datasets: the ones we have designated datasets 2 and 3, both produced using SELDI WCX2 chips. For each sample in the datasets, intensities are recorded at each of 15,154  $m/z$  values between 0 and roughly 20,000. Both datasets are available for download from <http://www.ncifdaproteomics.com>. Using these two datasets, we proceeded to carry out the following analyses:

Analysis 1: Is the proteomic pattern found by Zhu et al. biologically plausible?

The assumption behind mass spectrometry profiling for cancer screening is that some proteins will *consistently* be either more or less abundant in samples taken from cancer patients. To determine whether the intensities at the 18 m/z values identified by Zhu et al. (13) in December of 2003 (henceforth referred to as Zhu's values) appear to represent such a reproducible pattern of protein expression, we examined datasets 2 and 3 at each reported m/z value. After applying the processing steps that Zhu et al. describe (normalization and smoothing with a half-width Gaussian kernel), we calculated the mean value and standard deviation for the cancer and control samples at each m/z value in the pattern. The resulting distributions for the cancer and control samples were compared using simple two-sample T-statistics, computed separately for each dataset. The 16 non-cancer disease samples in dataset 2 were omitted. The sign of the resulting T-value indicates the direction of protein expression observed; if the T-value is positive, then the protein at that m/z value is expressed more abundantly in control patients. Negative T-values indicate that the protein in question is expressed more strongly in cancer patients. For each of the m/z values tested, we compared the T-statistics from the two datasets to check for gross differences in the pattern of protein expression at that site. The biologically plausible finding would be T-values whose magnitude and sign were roughly the same for each of the m/z values assessed in the two datasets.

Analysis 2: Is the near-perfect cancer/control separation obtained by Zhu et al. better than chance?

In addition to examining biological plausibility, we considered whether Zhu et al.'s findings demonstrate the presence of more separating structure than might be expected due to chance alone. Our hypothesis was that effective separation of cancer and control samples is actually quite common in dataset 3, when any randomly chosen 18 m/z values are used for classification. What is at stake is the assumption that Zhu's values from dataset 2 are more effective than any other set of 18 m/z values at classifying dataset 3 because of some common, biologically important structure. If randomly chosen values serve to discriminate cancer from control in dataset 3 as well as Zhu's values, then this assumption must be abandoned.

We addressed this question using four simulations, in which 18 randomly chosen m/z values were used to try to separate groups of spectra from dataset 3. In order for Zhu's values to be believable as a demonstration of reproducibility, they should perform significantly better than these baseline values. All groups to be separated reflected the ratio of cancer to control patients – 162 to 91 – that is seen in dataset 3. We used Matlab to perform all of the simulations; the code we employed is available at <http://bioinformatics.mdanderson.org>. We assessed the group membership using Fisher's Linear Discriminant Analysis (LDA): the 18 m/z values and 253 samples produced an 18 by 253 matrix of intensities, to which we applied LDA to find the direction in 18-space where the group patterns were most distinct when projected onto a single straight line. Looking only at the projections of the data onto this vector, we chose a cut point providing the best separation of the two groups. We then evaluated the set of samples found on each side of the cut point, and designated them as a group, based on the identity of the majority of the samples included on that side. Thus, if the two groups are

distributed completely randomly, there will be more members of the larger group on both sides of the cut point, and both sets will be predicted to be of the type of the larger group. After this labeling, we recorded the fraction of the samples (of 253) that were correctly classified. To account for random variability in the permutation-based estimates of accuracy, we repeated the above process 1000 times. The process resulted in a prediction of how accurately random  $m/z$  values can discriminate between two groups (such as cancer and control). We chose to use LDA for group separation because of its simplicity and speed, as thousands of simulations were to be run. LDA is somewhat simpler than the  $k$ -nearest-neighbors approach used by Zhu et al. (13); using their technique should produce even better classification results.

In the first simulation, we assessed how well 18 random  $m/z$  values could discriminate between two random, arbitrarily assigned groups. We note that in dataset 3, the simple rule of designating all samples “cancer” will be correct for 162/253 cases, yielding a minimum prediction accuracy of 64.09% for this dataset. In this first simulation, we employed two stages of randomization. First, we pooled the 253 samples in dataset 3 and chose 91 at random to comprise group A, assigning the remaining 162 to group B. Due to the random nature of the selection, there should be no systematic difference between the two groups. Second, we chose 18  $m/z$  values at random from the complete set of 15,154 available. We performed the assessment of group membership as described above, selecting a new set of  $m/z$  values and a fresh assignment of spectra into groups A and B for each of the 1000 simulations. In summary, the results of the first simulation set a baseline for how well 18 values can be used to separate two groups when no true structure exists.

For the second simulation, we assessed how well 18 random  $m/z$  values could separate the true cancers from the true controls in dataset 3; there is no random “partitioning into groups” step. As above, we chose 18  $m/z$  values at random from the complete set of 15,154 available, and used those values for the assessment of group membership (as described above). Again, we chose a new set of  $m/z$  values at random for each simulation. This second simulation, then, establishes a benchmark for how well the actual cancers and controls can be segregated using 18 random  $m/z$  values – a general benchmark to be used when deciding whether the segregation achieved by Zhu et al. is meaningful.

In two final simulations, we addressed whether random  $m/z$  values chosen from the lower end of the  $m/z$  range would be even more effective for group discrimination in dataset 3. With Zhu’s values, most of the separating structure in the data sets is to be found in the lower  $m/z$  range. If this is due to low molecular weight cancer-specific proteins in the sample, then random  $m/z$  values chosen in this range should not classify the samples well, compared to the values identified in Zhu et al. (13) On the other hand, if the lower  $m/z$  range is particularly affected by systematic bias in processing, due to differences in matrix batch or electronic noise effects that diminish in the higher  $m/z$  ranges, then randomly chosen sets of low-range  $m/z$  values should outperform sets chosen from the entire range and might better mimic the results observed by Zhu et al. Hence, we repeated the process described in simulation 2 twice more, but we restricted it by selecting only  $m/z$  values from the subsets of values (a) below 6000 (8302 indices), or (b) below 1000 (3390 indices).

Our approach has limitations. While our simulations may establish that we need not posit the existence of preserved biological structure to explain why  $m/z$  values selected to separate cancer from control in dataset 2 would work to separate cancer from control in dataset 3, they cannot rule out the existence of such structure. Thus, while our analyses may suggest that random chance is a plausible explanation for the ability of Zhu's values to separate cancer from control samples in dataset 3, we cannot contend that biological structure is not present, merely not clearly apparent.

## Results

### Analysis 1: Is the proteomic pattern found by Zhu et al. biologically plausible?

The two-sample t-statistics for all 18 of Zhu's values, from dataset 2 and dataset 3, are plotted in figure 1. 16 of the 18 t-values in dataset 2 are greater than 4.22 in absolute value, and 8 of 18 in dataset 3. However, the sign of the t-value differs between the datasets for 13 of 18 points. As noted, this leads to an implausible biologic explanation for the discriminatory ability: For over 2/3 of the  $m/z$  values, the measured protein appears to be expressed more in cancer samples for one dataset and more in controls for the other.

### Analysis 2: Is the near-perfect cancer/control separation obtained by Zhu et al. better than chance?

The results of all of our simulations are summarized in figure 2. As the figure demonstrates, the median accuracy of prediction obtained in the first simulation, where both sample groupings and  $m/z$  values were chosen at random, is about 70%. This is

somewhat above the baseline prediction of 64.09%, which is to be expected when multiple m/z values are used for classification (akin to the use of multiple comparisons). This finding illustrates the level of accuracy that would be required to prove that a methodically selected set of m/z values is meaningfully predicting group membership in general.

When the supplied cancer/control labeling is used in the second simulation, the median prediction accuracy is slightly better than 96%, with a small percentage of simulations achieving perfect accuracy. As shown in the final two simulations in figure 2, the cancer and control groups are even more easily classified using 18 randomly-chosen low-range m/z values. When all 18 values are below 6000, the median prediction accuracy is better than 97%. When all 18 values are below 1000, the median prediction accuracy climbs to about 99%, and in more than 20% of trials perfect classification is achieved. In summary, these randomly chosen values perform as well as Zhu's values for classification of dataset 3, suggesting that we do not need to posit the existence of any special biological structure associated with the proteomic pattern to explain the results.

## Conclusions

In the preceding pages, we have demonstrated that the proteomic pattern derived from dataset 2 is not biologically plausible when applied to dataset 3. This pattern fails to demonstrate either biological plausibility or reproducibility for the ovarian cancer datasets.

The fact that extremely good separation between cancers and controls is achieved when m/z values are chosen at random suggests that it is not necessary to posit the

existence of a preserved biological structure between datasets 2 and 3 to explain the classification rates observed by Zhu et al. (13) In fact, it seems quite unlikely that the cancer samples differ from controls in the expression of every protein measured. To us this suggests the presence of systematic bias, in the presence of which dataset 3 cannot be realistically used for confirmation of findings from any other ovarian cancer spectra.

## Discussion

Ultimately, this paper raises several questions: First, why are the ovarian cancer datasets so remarkably classifiable, given that this classification does not seem to reflect true biological differences? Second, can the data contained in ovarian datasets 2 and 3 be used for clinical applications? Third, can these problems be avoided using a higher precision instrument? Fourth, what lessons can be learned from these datasets to guide the design of future experiments using mass spectrometry for cancer detection? Finally, do our findings mean that proteomic profiling in general is not worth pursuing?

### Why are the ovarian cancer data sets so remarkably classifiable?

With respect to the extreme separation noted for dataset 3, the pervasive nature of the separation again suggests experimental bias - in keeping with previous findings (11, 12). Bias can introduce artificial separation, and once these artificial factors have been introduced, they are “hard-wired” and difficult if not impossible to remove. The exact nature of the bias is unclear. It could involve running all samples of one group on separate chips, running them at a different time than the other spectra if the machine is subject to drift, or using different chemical batches of matrix. All of these implicitly

assume that the cancers and controls were not randomly intermingled on the same chips, as such treatment would render the biases mentioned above extremely difficult to introduce. It is also possible to introduce separating bias by using different collection protocols for different groups of samples, though we think that is unlikely here.

More recently, we have observed that the filenames originally associated with the spectra in dataset 3 suggest one source of experimental bias. Dataset 3 was originally posted in June of 2002 at <http://clinicalproteomics.steem.com>, and was later reposted in August of 2002 with the names of the spectra changed (the new filenames match those from some higher resolution scans; see below). The earlier version of dataset 3 is no longer publicly available. In the version of dataset 3 posted in August of 2002, the filenames of individual spectra are cryptic. However, a typical control sample from the dataset as posted in June is named as follows:

wcx2 control d 382-ca602-wcx2-d

Cancer sample files are named similarly:

wcx2 ovarian f 431-cb481-wcx2-a

Figures 4 and 5 demonstrate our interpretation of the way in which these file names reveal the chip type, chip identification, spot number, and sample type for the samples in this dataset, and how this information is supported by the posted Excel file giving clinical information for the cancer samples in dataset 3. In addition, each filename contains a 3-digit number which is, without exception, associated with exactly one unique chip ID. In the control samples, these 3-digit numbers form a contiguous sequence from 382 to 405; in the cancer samples, the corresponding numbers range from 430 to 465. There is no overlap between the cancer and control designations for either

the chip ID or the 3-digit portion of the filename. This suggests that the cancer samples were all run on different chips than the control samples. This lack of randomization may explain the remarkable separation of cancer and control samples in this dataset, either through biases associated with the chips themselves, or through the use of different chips allowing for the possibility of other types of bias such as running all of the control samples before all of the cancers.

Our identification of this source of bias rests on our interpretation of the file naming conventions used by the NCI/FDA clinical proteomics labs, specifically on the assumption that the chip IDs were part of the file names in the June 2002 posting. If this assumption is incorrect, it should be possible to provide the correct interpretation of the file names.

Can these datasets be used for clinical applications yet?

In our opinion, the answer is no, for two reasons. First, the effects of experimental bias in this data are of sufficient magnitude to completely obscure any biologically relevant information, particularly in dataset 3. Second, even if we were to assume that this data contains clinically useful information measurable despite the noise, any such assumptions should be experimentally reproduced before being tested in a clinical arena. Our examination of Zhu's values convinces us that they do not provide a demonstration of biological reproducibility, and no other demonstration has been provided to date.

Can these problems be avoided by using a higher precision machine?

Unfortunately, the answer to this question is also no. A higher precision machine would help resolve the problem if the cancers and controls had been processed in an identical fashion, so that the only source of consistent differences should be the underlying biology. In that case, improving the resolution would let us pick up smaller differences, and we could trust that they were real. The problems that we have noted are associated with the introduction of systematic biases through imperfect randomization. A higher-resolution machine will simply record both effects, bias and biology, with greater accuracy – without resolving the risk of mistaking one for the other. An example of this problem can be seen in the ovarian cancer Qstar data experiment described in Conrads et al. (17). In this experiment, 248 spectra were initially produced: 152 cancer and 96 control. A QA/QC filter was then applied to remove “lesser quality” spectra. As noted in Conrads et al. (17), the lesser quality “mass spectra were all generated at the end of the experimental run,” which is important as they note that the machine was in the process of breaking down and spectra run later were more affected. However, of the 32 spectra identified as being of “lesser quality”, 31 are cancers. If the cancers and controls have been randomly commingled, the two-sided p-value associated with this result is  $9.98e-07$ , literally “one in a million”. This observation is much easier to explain if we assume that cancers and controls were run on different chips, as we believe was the case with SELDI dataset 3. We note in passing that the filenames supplied for SELDI dataset 3 in August of 2002 coincide with the filenames of the Qstar spectra, which were also derived from SELDI chips. If the SELDI chips from dataset 3 were used in the Qstar experiment, then biases affecting the dataset 3 chips are directly relevant. While some biological structure

is likely present, and indeed easier to detect with a higher resolution instrument, there can still be bias.

What can we learn from the NCI/FDA study to help in designing future experiments?

First and foremost, it is clear that sample randomization is critical to reduce bias. This randomization must happen at the level of chip, chip spot, batch order, and run order. Stringent randomization and calibration will be critical for the success of any biological mass spectrometry experiment. Second, statistical analysis of spectra should always be accompanied by a sanity check: Are the patterns of protein expression consistent across multiple datasets? Ideally, this check should be performed on a new dataset collected and run either later or elsewhere. Finally, spectra analysis should always be validated by confirming that the separation seen with the chosen (and presumably biologically relevant)  $m/z$  pattern is a meaningful improvement over that found with randomly chosen  $m/z$  values.

The NCI/FDA group is to be commended for making large portions of their data publicly available. While our analysis has involved a certain amount of extrapolation and detective work, it would have been impossible without the published data. We are well aware that the experimental biases measurable in the ovarian cancer datasets may affect other groups' efforts at protein profiling as well. Therefore, we urge other groups to match or exceed the NCI/FDA policy of publishing supporting data for mass spectrometry results. In addition to the raw spectra, other information which should be made available for peer review includes calibration settings and spectra, the

randomization protocol, and the ultimate chip spot, batch position, and run order for each sample.

Do our findings mean that proteomic profiling in general is not worth pursuing?

No, our findings do not invalidate proteomic profiling in general. The problems we have discussed here, once recognized, can be dealt with effectively. This requires careful experimental design, with conscientious randomization. Proteomic profiling is still very much in development. The very sensitivity that gives these assays their predictive potential can also render them misleading if biases are present. Addressing these issues will be the next step to exploiting the potential of this promising approach to cancer detection.

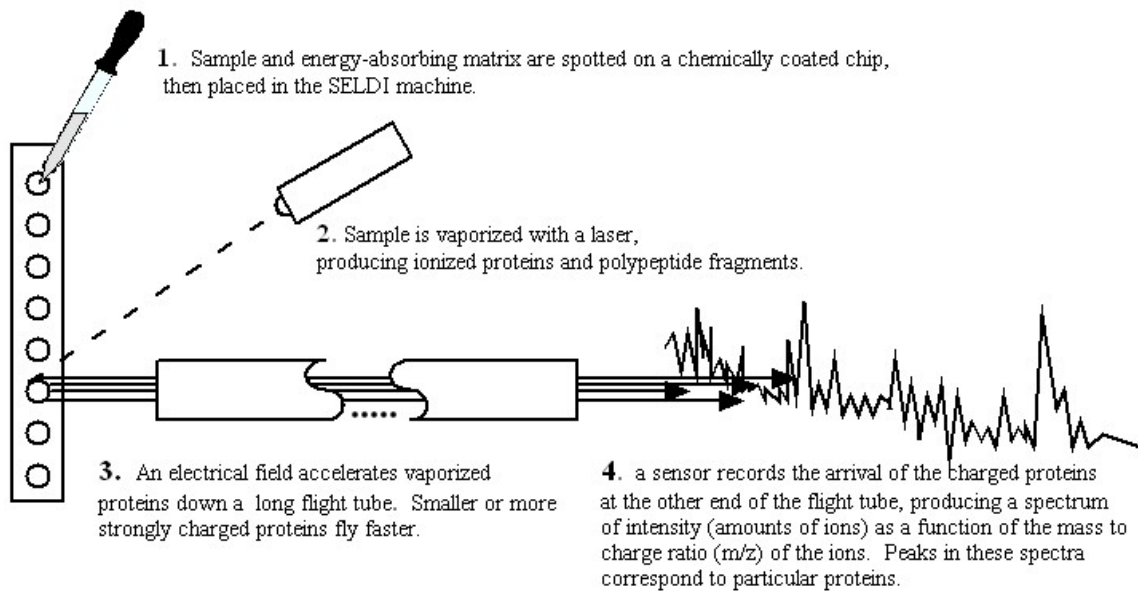
## References

1. Pecorelli S, Favalli G, Zigliani L, Odicino F. Cancer in women. *Int J Gynaecol Obstet* 2003;82(3):369-79.
2. 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(9306):572-7.
3. Adam B-L, Qu Y, Davies JW, Ward MD, Clements MA, Cazares LH, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* 2002;62:3609-3614.
4. Koopmann J, Zhang Z, White N, Rosenzweig J, Fedarko N, Jagannath S, et al. Serum Diagnosis of Pancreatic Adenocarcinoma Using Surface-Enhanced Laser Desorption and Ionization Mass Spectrometry. *Clinical Cancer Research* 2004;10:860-868.

5. Li J, Zhan Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry* 2002;48:1296-1304.
6. Petricoin EF, 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 2002;94(20):1576-8.
7. Rai AJ, Zhang Z, Rosenzweig J, Shih Ie M, Pham T, Fung ET, et al. Proteomic approaches to tumor marker discovery. *Arch Pathol Lab Med* 2002;126(12):1518-26.
8. Ye B, Cramer DW, Skates SJ, Gygi SP, Pratomo V, Fu L, et al. Haptoglobin-alpha subunit as potential serum biomarker in ovarian cancer: identification and characterization using proteomic profiling and mass spectrometry. *Clin Cancer Res* 2003;9(8):2904-11.
9. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* 2004;3(4):367-78.
10. 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(20):6971-83.
11. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF Protein Patterns in Serum: Comparing Data Sets from Different Experiments. *Bioinformatics* 2004;20:777-785.
12. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;4(1):24.
13. 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(25):14666-71.
14. Petricoin EF, 3rd, Fishman D, Conrads T, Veenstra T, Liotta L. Proteomic pattern diagnostics: producers and consumers in the era of correlative science. *BMC Bioinformatics*, comment posted March 2004. <http://www.biomedcentral.com/1471-2105/4/24/comments>.

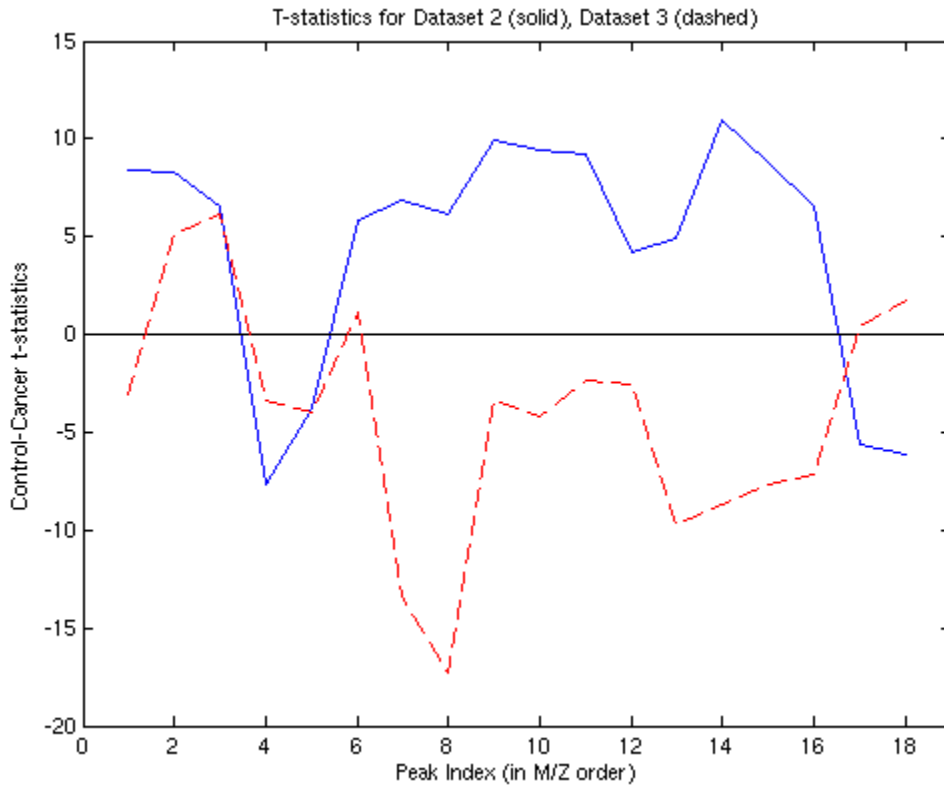
15. Graham R. Growing plasma proteome sets stage for new diagnostics. *J Proteome Res* 2004;3(2):163.
16. Check E. Running before we can walk? *Nature* 2004;429(6991):496-497.
17. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, et al. High-resolution serum proteomic features for ovarian cancer detection. *Endocr Relat Cancer* 2004;11(2):163-78.
18. Clinical Proteomics Program Frequently Asked Questions. Available at <http://www.ncifdaproteomics.com/faq.doc> Viewed 4/22/2004.

Figure 1:



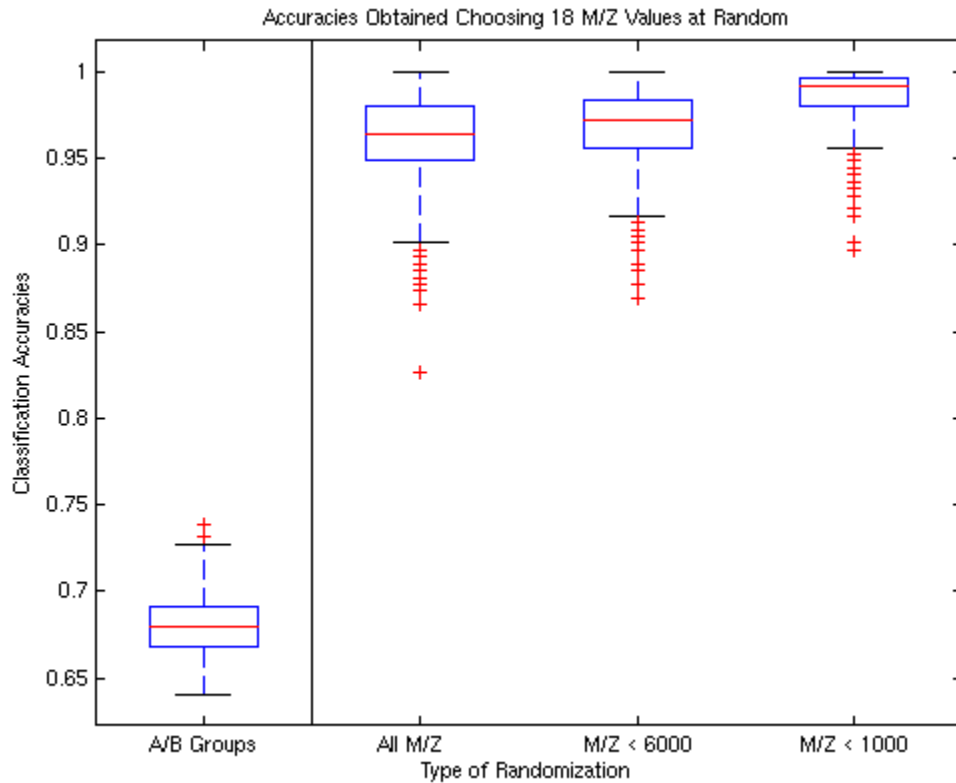
Description of the mass spectrometry variant known as surface-enhanced laser desorption and ionization (SELDI): (1) A preparation of a biological sample such as serum is applied to a chemically coated chip. A subset of the proteins contained in the sample bind to the chip, which, after the addition of an energy absorbing matrix, is inserted in the SELDI machine. Next, (2) vaporization of the bound proteins from the matrix crystals with a laser produces ionized proteins or polypeptides (which we will simply call proteins in this paper), which are then (3) accelerated down a flight tube by an electric field. Measurements (4) of the time of flight for the different proteins traveling down the tube are used to assemble a spectrum of intensity (amounts of ions) as a function of the mass to charge ratio ( $m/z$ ) of the ions.

Figure 2:



Reanalysis of discriminating  $m/z$  values identified by Zhu et al. (13) T-statistics for each value show two pieces of information: whether the difference between cancers and controls at an  $m/z$  value is statistically significant (magnitude greater than about 4), and what direction that difference is in (positive values suggest higher expression levels in controls). The t-statistics for datasets 2 (solid line) and 3 (dashed line) suggest that the proteomic pattern found by Zhu et al. does not have a plausible biological explanation. While the t-statistics do show significant differences between cancers and controls at most of the  $m/z$  values identified, expression levels at 13 of the 18 peaks were higher in cancers for one dataset and higher in controls for the other.

Figure 3



Results of simulations testing discriminatory ability of random m/z values in dataset 3.

Simulations from left to right include:

- A/B Groups: Expected classification accuracy for 18 random m/z values when groups are chosen at random (null hypothesis);
- All M/Z: Expected classification accuracy for 18 random m/z values when original cancer and control groups are classified;
- M/Z < 6000: Expected classification accuracy for 18 random m/z values below 6000 when original cancer and control groups are classified;
- M/Z < 1000: Expected classification accuracy for 18 random m/z values below 1000 when original cancer and control groups are classified.

The results demonstrate that classification of dataset 3 is extremely accurate even when random values are employed, thus calling into question the significance of equivalent classification accuracy demonstrated by Zhu's values.

Figure 4:



A photograph of part of a WCX2 chip from 2002. Labeling of the spots is apparent, as is the form of the etched ID: two letters followed by 3 numbers.

Figure 5:

8-7-02 SELDI Chip ID	6-19-02 File Names	
430-CB533-WCX2- <b>C</b>	wxc2 ovarian e 430-cb533-wxc2- <b>c</b>	<b>Spot on chip</b>
430-CB533- <b>WCX2</b> -D	wxc2 ovarian f 430-cb533- <b>wcx2</b> -d	<b>Chip type</b>
430- <b>CB533</b> -WCX2-F	wxc2 ovarian f 430- <b>cb533</b> -wxc2-f	<b>Unique Chip ID</b>
<b>430</b> -CB533-WCX2-H	wxc2 ovarian f <b>430</b> -cb533-wxc2-h	<b>3-digit number corresponds exactly to Unique Chip ID</b>
	wxc2 <b>control</b> d 382-ca602-wxc2-d	<b>Designates cancer or control</b>

Demonstration of parallel structure between sample information contained in the Excel file included with the second posting of dataset 3 (shown in first column) and file names used in the initial posting of dataset 3 (shown in second column). As shown, file names appear to contain information about chip type, chip identification, and the spot used on the chip for a given sample.

Table 1: SELDI Ovarian Cancer data made available by the NCI/FDA Clinical Proteomics Program

Dataset*	When posted	Chip type	Subjects tested	Used for training set	results
Dataset 1	2/16/2002	H4**	100 control 100 cancer 16 benign disease	50 cancer 50 control	50/50 cancer samples called cancer 46/50 control samples called control 4/50 control samples called cancer 16/16 benign disease classed as "other" 5 m/z values used in pattern
Dataset 2	4/3/2002	WCX2	Same as dataset 1	50 cancer 50 control	"improved" from dataset 1 5 m/z values used in pattern
Dataset 3	8/7/2002***	WCX2	91 Control 162 Cancer	Not reported	100% correct classification; 7 m/z values used in pattern

\*All three datasets are available for download at <http://www.ncifdaproteomics.com>

\*\*Because the H4 chip has now been discontinued, it is not possible to pursue further efforts to develop or replicate these results(18).

\*\*\*Originally posted in June 2002, at <http://clinicalproteomics.stem.com>, this file was re-posted in August 2002 with different file names for the individual spectra.

Table 2: Summary of some publications noting problems with the mass spectrometry protein profiling technique and the ovarian cancer datasets.

Paper	Key points
Diamandis (9)	<ul style="list-style-type: none"> <li>• The SELDI technique can only identify proteins that are abundant in the blood; proteins associated with small tumors will be too scarce to be seen with this technique.</li> <li>• Many proteins seen will be nonspecific cellular reactants rather than cancer-specific proteins.</li> <li>• Many of the proteins which have been identified have previously been examined and were not helpful for diagnosis</li> <li>• There have been difficulties with proving the approach to be reproducible over time.</li> </ul>
Rogers et al. (10)	<ul style="list-style-type: none"> <li>• Highly diagnostic SELDI protein patterns in renal cancer proved far less effective for data analyzed several months later.</li> </ul>
Sorace and Zhan (12)	<ul style="list-style-type: none"> <li>• In dataset 3, it is possible to achieve perfect separation using regions of the spectra which are normally thought to be unstable and driven by electronic noise.</li> </ul>
Baggerly et al. (11)	<ul style="list-style-type: none"> <li>• Dataset 1 shows clear evidence that benign disease samples were handled systematically differently from other samples.</li> <li>• All three datasets show evidence of inconsistent processing.</li> <li>• All three datasets show evidence of incorrect calibration, which would make it impossible to compare datasets and demonstrate reproducibility of the results.</li> </ul>