A Cautionary Note on the Evaluation of Biomarkers of Subtypes of a Single Disease

Adeniyi J. Adewale¹, Qi Liu², Irina Dinu², Paul D. Lampe³, Bree L. Mitchell⁴, and Yutaka Yasui²

¹ Merck Research Laboratories, North Wales, PA.
² Department of Public Health Sciences, School of Public Health, University of Alberta, Edmonton, Alberta, Canada.
³ Fred Hutchinson Cancer Research Center, Seattle, WA.
⁴ Biomedical Life Sciences Group, Intel Corporation, Santa Clara, CA.

Received for publication December 7, 2007; accepted for publication May 22, 2008.

Heterogeneity in the molecular characteristics of a disease presents a challenge to investigators attempting to identify biomarkers of the disease. Preceding the biomarker discovery effort with stratification within a heterogeneous disease group, which amounts to grouping disease cases into more homogeneous subtypes, seems to be a natural strategy for discovering subtype-specific biomarkers. This is because biologically more homogeneous subgroups are presumably easier to distinguish from the nondiseased than the entire heterogeneous disease group. The misleading benefits of this two-step approach are illustrated using an example from a protein biomarker discovery project for breast cancer. A potential analytical pitfall in this framework is explained using a conditional probability argument.

biological markers; classification problem; conditional probability; cross-validation; mass spectrometry; misclassification

Abbreviation: MALDI-TOF, matrix-assisted laser desorption/ionization time-of-flight.

A disease can be highly heterogeneous in its molecular characteristics, even if it is labeled as a single disease taxonomically. Heterogeneity within a disease presents a serious challenge to investigators attempting to identify biomarkers that can be used for early detection and tailored treatment. A preliminary class-discovery analysis leading to subtyping within the disease is a natural strategy for biomarker discovery for a seemingly heterogeneous disease. For early detection of a disease, for example, it is plausible that biologically more homogeneous subgroups are easier to distinguish from the nondiseased than the entire heterogeneous disease group.

In this paper, we submit a cautionary note on evaluating biomarkers of such subtypes of a single disease, taking a detection (classification) problem of disease cases from disease-free controls. To illustrate the point concretely, we describe a subtyping and classification analysis based on matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) plasma protein expression in breast cancer cases and breast-cancer-free subjects. The classification results following subtyping of the breast cancer patients appeared substantially better than those obtained without subtyping. We explain a fallacy presented by the apparently improved classification following the subtyping, using simple conditional probability arguments. We conclude by discussing the implications of heterogeneity problems in biomarker discovery.

ILLUSTRATION WITH MALDI-TOF BREAST CANCER CASE-CONTROL STUDY

Breast cancer is an example of a disease with heterogeneity in its molecular characteristics (1). To illustrate the
fallacy, we consider an exploratory biomarker-discovery project in which breast cancer cases and controls were compared with respect to their prediagnostic expression patterns of plasma proteins as measured by MALDI-TOF mass spectrometry techniques. Briefly, the project involved 198 breast cancer cases and 198 randomly selected controls who were frequency-matched on age to the cases. All of the subjects were participants in a large epidemiologic cohort study at the Aichi Cancer Center (Nagoya, Japan) in which blood samples were collected, processed, and stored prior to the diagnosis of cancer. At the Fred Hutchinson Cancer Research Center (Seattle, Washington), MALDI-TOF mass spectrometry was used to measure protein expression in each of the 198 cases and 198 controls, using the protocol established previously (2). After preanalysis preprocessing of the MALDI-TOF mass spectra data (3), the data consisted of 2,319 protein expression levels, each of which is a potential biomarker with which the cases and controls can be distinguished.

In an attempt to identify a panel of protein biomarkers of breast cancer, we employed boosting, a popular high-dimensional classification method, to form a panel of proteins for classification of breast cancer cases and controls (4–7). Fivefold cross-validation was used to estimate the error of the boosting classification. A minimum cross-validation error of 44.2 percent was obtained using 13 proteins in the classification panel. This observed classification error is slightly better than random guessing, which would result in a classification error of 50 percent. This underscores the difficulty of identifying breast cancer biomarkers.

We hypothesized that this difficulty was due, at least partly, to the known heterogeneity within breast cancer. Therefore, subtyping of the breast cancer cases using their MALDI-TOF protein expression data was considered before application of the boosting classification method. We employed principal-components analysis to reduce the dimensionality of the data, in which 182 principal components with eigenvalues greater than or equal to 1 were retained. They were used in a subsequent k-mean cluster analysis with the desired number of clusters specified as five. Thus, five groups ("subtypes") of breast cancer were created, each of which was more "homogeneous" in protein expression than the entire group of breast cancer cases. The numbers of cases in the five subtype groups were 28, 34, 37, 37, and 62.

Our strategy of subtyping was to stratify the breast cancer cases into more homogeneous subtypes and discover more discriminatory biomarkers for each subtype against the controls. In seeking biomarkers of subtypes of breast cancer, we repeated the boosting classification analysis, comparing each subtype with the 198 controls. The minimum cross-validation errors attained for the subtypes with 28, 34, 37, 37, and 62 cases were 12.4 percent, 14.7 percent, 15.7 percent, 15.7 percent, and 23.8 percent, respectively (table 1).

There is an apparently substantial gain obtained by subtyping the breast cancer cases prior to the boosting classification, since all subtypes have reduced cross-validation errors significantly to less than 25 percent, compared with the overall classification error of 44.2 percent in the original whole group classification. Below, we explain why this apparently substantial gain obtained by subtyping is not a real gain and why this is a fallacy introduced by the subtyping and the classification of each subtype versus the controls.

**FALLACY OF RESULTS EXPLAINED**

We clarify the fallacy as follows. Define the sets $D_k$ and $C$ as the groups of subjects with disease subtype $k$ and control subjects, respectively, where the union of all subtypes, $D = \cup D_k$, is the entire disease group ($D_k$’s are disjoint and the subtypes are mutually exclusive). Let $|A|$ denote the number of members in set $A$ (the cardinality of $A$). The classification error in the classification analysis of $D_k$ versus $C$ is an estimate of $p(\text{misclassification}|D_k \text{ or } C)$, which can be expressed as follows:

$$p(\text{misclassification}|D_k \text{ or } C) = p(\text{call } C|D_k) \times (p(D_k|D_k) \text{ or } C) + p(\text{call } D_k|C) \times p(C|D_k) \text{ or } C$$

$$= p(\text{call } C|D_k) \times |D_k|/(|D_k| + |C|) + p(\text{call } D_k|C) \times |C|/(|D_k| + |C|).$$

The specific components of the overall misclassification probability are the false-negative probability of calling a control given a disease subtype $k$ subject, $p(\text{call } C|D_k)$, and the false-positive probability of calling a disease subtype $k$ given a control subject, $p(\text{call } D_k|C)$. Now, consider a hypothetical classifier that classifies all subjects as controls such that $p(\text{call } C|D_k) = 1$ and $p(\text{call } D_k|C) = 0$. Clearly, this hypothetical classifier is useless, and yet $p(\text{misclassification}|D_k \text{ or } C) = |D_k|/(|D_k| + |C|)$, which could be substantially smaller than $|D|/(|D| + |C|)$, the classification error of this hypothetical classifier in the overall classification of $D$ versus $C$.

Thus, a classification error in the magnitude of $|D_k|/(|D_k| + |C|)$ in the classification of disease subtype $k$ versus controls does not necessarily translate into a good classifier, no matter how small $|D_k|/(|D_k| + |C|)$ is. This is because $|D_k|/(|D_k| + |C|)$ can be made as small as desired by labeling a small subset of the cases as subtype $k$, while the classifier remains useless. Thus, the drastic reduction in classification errors following subtyping when compared with the overall classification error of the entire disease group versus controls in the MALDI-TOF breast cancer case-control data should not be interpreted as improved classification performance.

The above argument establishes that the reduced misclassification error rate in every subtype classification against controls does not imply that real biomarkers of breast cancer subtypes have been identified. However, the argument does not prove that the resulting subtype classifiers are useless. In order to assess the utility of the subtyping approach, combined measures of performance across subtypes are required. The combined overall misclassification probability following subtyping, that is, $p(\text{misclassification})$, can be regarded as an unconditional error rate, while the subtype-specific misclassification probability, $p(\text{misclassification}|D_k \text{ or } C)$, can be regarded as a conditional error rate. The cross-validated subtype classification error is an estimate of the
conditional probability of misclassification, given that the sample to be classified is either a case of the specific disease subtype or a control—definitely not a case of the other subtypes of the disease. This is not a realistic condition that is met in real classification problems. For a subject to be classified, it is unknown whether the subject has the disease or not. To know that the subject either has a specific subtype of the disease or is disease-free, one needs to know that the subject does not have one of the other subtypes of the disease. Thus, the reported subtype-specific conditional probability of misclassification is not a realistic quantity of importance in practice. However, the combined overall misclassification probability can be estimated as follows:

\[ p(\text{misclassification}) = p(D)(1 - \text{sensitivity}) + (1 - p(D))(1 - \text{specificity}), \]

where the sensitivity and specificity are the combined sensitivity and combined specificity, respectively, following subtyping. The combined sensitivity and specificity following subtyping can be expressed in terms of estimable probabilities from each subtype classification:

\[ \text{Sensitivity} = p(\text{call } D|D) = 1 - p(\text{call } C|D) = 1 - \sum_{k=1}^{m} p(\text{call } C|D_k)p(D_k|D) \]

\[ = 1 - \sum_{k=1}^{m} \frac{|D_k|}{|D|} p(\text{call } C|D_k), \]

\[ \text{Specificity} = 1 - p(\text{call } D|C) = 1 - \sum_{k=1}^{m} p(\text{call } D_k|C). \]

Table 1 presents the estimates of the overall misclassification error, sensitivity, and specificity from the classification without subtyping and with subtyping. The combined overall misclassification error following subtyping is 50.0 percent. The combined sensitivity and specificity following subtyping are 0.5 percent and 99.5 percent, respectively, compared with 58.6 percent and 53.0 percent without subtyping. The combined assessment following subtyping clearly reveals that there is no real benefit from the subtyping. The combined sensitivity and specificity following subtyping show that the subtyping approach—preceding classification with class discovery—amounts to the simple trading of sensitivity for specificity.

A pragmatic approach to assessing the overall performance of the classification following subtyping is to devise an appropriate combination rule for utilizing the five subtype-specific classifiers. Note that the above overall performance calculation following subtyping still assumes the unrealistic ability of correctly identifying the disease subtype of a subject to be classified. A practical and intuitively more appealing rule is “believe any positive classification”—that is, classifying a presenting subject as a “case” if at least one of the five classifiers classifies the subject as a “case.” The “believe any positive classification” rule increases the overall sensitivity relative to the subtype-specific sensitivities, while at the same time it reduces the overall specificity. Using this rule, the combined overall misclassification rate, sensitivity, and specificity estimated by fivefold cross-validation were 49.0 percent, 1.5 percent, and 96.5 percent, respectively. Again, this provides evidence against the reduced subtype-specific misclassification errors.

**DISCUSSION**

When we seek subtype-specific biomarkers, caution must be exercised in evaluating their performance. Specifically, in each subtype classification, the classification error may appear lower than that in the overall disease-control classification. This may appear to support the advantage of considering subtypes over considering the disease as a single entity. However, this can be a fallacy.

In general, the overall misclassification error as a measure of classification accuracy can be misleading in the presence of imbalance in class sizes. The overall misclassification error can be as small as the ratio of the smaller class size to the total sample size, while the classifier may in fact be useless. Other measures of classification accuracy, particularly sensitivity and specificity, are desirable since they are not susceptible to distortion by relative class sizes (8). Had the sensitivity and specificity been considered in each subtype classification in our example, it would have been clear

---

**TABLE 1. Comparing the performance of classification with subtyping to that of classification without subtyping using sensitivity and specificity**

<table>
<thead>
<tr>
<th>Subtype vs. Control</th>
<th>No. of Disease Cases (Dk)</th>
<th>Overall Error (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtype 1 vs. control</td>
<td>28</td>
<td>12.4</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Subtype 2 vs. control</td>
<td>34</td>
<td>14.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Subtype 3 vs. control</td>
<td>37</td>
<td>15.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Subtype 4 vs. control</td>
<td>37</td>
<td>15.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Subtype 5 vs. control</td>
<td>62</td>
<td>23.8</td>
<td>1.6</td>
<td>99.5</td>
</tr>
<tr>
<td>Overall (following subtyping)</td>
<td>198</td>
<td>50.0</td>
<td>0.5</td>
<td>99.5</td>
</tr>
<tr>
<td>Overall (without subtyping)</td>
<td>198</td>
<td>44.2</td>
<td>58.6</td>
<td>53.0</td>
</tr>
</tbody>
</table>
that there was no real benefit from the subtyping, since each of the subtype classifiers had approximately 0 percent sensitivity and 100 percent specificity (table 1). This 0 percent sensitivity and 100 percent specificity from each subtype classifier imply that each is essentially a "label-all-control" classifier like the useless hypothetical classifier. The combined sensitivity and specificity, obtained either mathematically by a combination of the subtype sensitivity and specificity values using probability arguments or by a more pragmatic "believe any positive classification" rule, which are again near 0 percent (sensitivity) and 100 percent (specificity), respectively, underscore the fact that the overall effect of classification following subtyping in the entire sample is akin to labeling all presenting subjects as controls. This fact was masked by the subtype-specific overall misclassification rate.

Finally, the illustration presented here should not be interpreted as a failure of subtyping as a potentially useful approach to the problem of unknown heterogeneity of disease in biomarker discovery. Failure to identify useful biomarkers either with or without subtyping is potentially attributable to many causes, including a poor choice of classification methods, measurement errors, and the lack of distinguishing molecular characteristics between disease subjects and nondisease subjects. In particular, the importance of the accuracy and reliability of high-dimensional experimental data from new biotechnologies such as MALDI-TOF cannot be overemphasized. These are important issues in biomarker discovery analysis of high-dimensional data in general: they are not specific to analyses either with or without subtyping. Our central point was that, because the results following subtyping showed superficial benefits which were dismissed only after careful analytical scrutiny, caution must be exercised in evaluating the performance of such subtype-stratified approaches. The fallacy occurred because of two factors: 1) the imbalance of sample sizes between each subtype of cases and controls and 2) the use of overall classification error instead of sensitivity and specificity. The imbalance of sample sizes could occur frequently if the subtype classification were considered within a study in which the sample sizes of cases and controls were balanced or had a ratio of 1:m, where m is greater than 1. Caution must be exercised in such cases.

ACKNOWLEDGMENTS

This work was partly funded by the Alberta Heritage Foundation for Medical Research (through postdoctoral fellowships to A. J. A. and I. D. and a senior health investigator award to Y. Y.) and the Canada Research Chair Program, Canadian Institute of Health Research (Y. Y.).

Conflict of interest: none declared.

REFERENCES


Am J Epidemiol 2008;168:559–562