Breast Cancer (2009) 16:248–251
DOI 10.1007/s12282-009-0141-9

CONFERENCE PAPER

Presidential symposium: From standardization to personalization in breast cancer treatment

Algorithm for in vitro diagnostic multivariate index assay

Kikuya Kato

Received: 16 December 2008 / Accepted: 27 May 2009 / Published online: 8 July 2009

Abstract Recently emerging diagnostic tools such as MammaPrint and oncotype-DX are beginning to have impact on clinical practice of breast cancer. They are based on gene expression profiling, i.e., gene expression analysis of a large number of genes. Their unique characteristic is the use of a score calculated from expression values of a number of genes, for which the Food and Drug Administration (FDA) created a new diagnostic category entitled "in vitro diagnostic multivariate index assay (IVDMIA)." In contrast to conventional biomarkers, IVDMIA requires an algorithm to calculate the diagnostic score. The linear classifier is the preferred algorithm. When the number of diagnostic genes is n, each tumor is represented by a point in an n-dimensional space made from gene expression values. Diagnostic algorithms (linear classifier) make an $(n-1)$-dimensional plane in the n-dimensional space to separate two patient groups. Calculation of the diagnostic score is achieved by dimension reduction. Currently, IVDMIA is restricted to gene expression profiling, and will also be applied to malignancies other than breast cancer.

Keywords Gene expression profiling · Prognostic factor · Linear classifier

This article is based on a presentation delivered at Presidential Symposium, "From standardization to personalization in breast cancer treatment," held on 26 September 2008 at the 16th Annual Meeting of the Japanese Breast Cancer Society in Osaka.

K. Kato (✉)
Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-3 Nakamichi, Higashinari-ku, Osaka 537-8511, Japan
e-mail: katou-ki@mc.pref.osaka.jp

Introduction

Recently emerging diagnostic tools such as MammaPrint [1, 2] and oncotype-DX [3] are beginning to have impact on clinical practice of breast cancer. Both should be considered in the context of personalized medicine for determination of chemotherapy through genetic information. They are based on gene expression profiling, i.e., gene expression analysis of a large number of genes. The unique characteristic of these tools is the use of a score calculated from expression values of a number of genes. This is a new feature that no previous diagnostic procedure has, for which the FDA created a new diagnostic category entitled "in vitro diagnostic multivariate index assay (IVDMIA)." Currently, MammaPrint and Pathwork of Origin Test (a microarray-based diagnostic system to determine tissue origin of cancer whose tissue origin is unknown) have been cleared by the FDA.

So far, when using a biomarker, for example, for estrogen receptor or blood cholesterol, the concentration or amount of the molecule is used as a score for diagnosis. However, in IVDMIA, the score is calculated from a number of measurement values, which are gene expression values in the cases of MammaPrint and oncotype-DX. Thus, in IVDMIA, the algorithm, i.e., the method of calculating the score, is critical. However, such an algorithm is a "black box" for clinicians. In this short review, I present a simplified explanation of the algorithm for IVDMIA.

Overview of the IVDMIA diagnostic system

There are two types of statistical analysis for gene expression profiling: unsupervised analysis and supervised

prediction [4, 5]. For diagnostic purposes, supervised prediction has usually been adopted. In supervised analysis, parameters of a diagnostic algorithm are determined with a learning set, and the performance of the algorithm is evaluated with a test set. With a small sample set, cross-validation procedures are usually applied. Leave-one-out (LOO) cross-validation is the most frequently used. In LOO, one sample is withdrawn, and the diagnostic classifier is built with the rest of the samples. The performance of the classifier is evaluated on the withdrawn sample. Repeating with all the samples, the overall performance of the classifier is determined. A schematic representation of LOO is shown in Fig. 1.

Although there have been many studies on supervised prediction, there have been very few comparing algorithms. The main obstacle is multiplicity of statistical test. Statistical significance should be adjusted when the test is repeated. For example, a prize could be obtained easily by increasing the number of lots drawn. The chance estimation of the prize should be adjusted with the number of trials. Similarly, when a number of classifiers are tested, some classifiers yield a good performance by chance. In addition, such studies do not guarantee consistency of results with other data sets. It should also be noted that many studies lacked proper evaluation of classifiers [4]. Thus, it is extremely difficult to determine the real performance of a classifier.

Although only a few comparative studies have been reported, a trend in choice of algorithm has been established. Relatively simple algorithms, categorized as linear classifiers, such as weighted voting [6] and nearest centroid [7], are now preferred. Complex algorithms such as artificial neural network [8] are the minority. One reason is that linear classifiers have sufficient performance. The other reason is that it is difficult to control overfitting with complex algorithms. Overfitting is a phenomenon inherent to supervised prediction: parameters of any algorithm are optimized with the learning data set, and its performance with other data sets is usually not as good as that with the learning set. In general, prevention of overfitting is easier

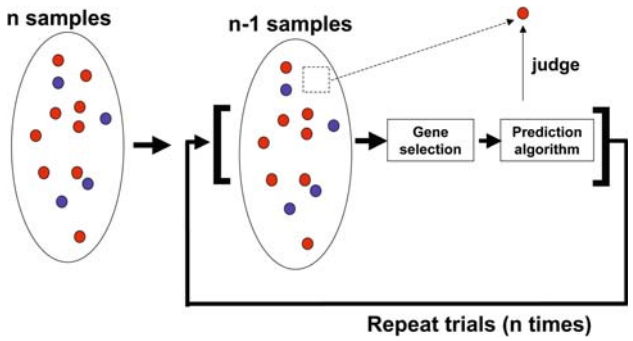

Fig. 1 Schematic representation of leave-one-out validation

with linear classifiers. The algorithm for MammaPrint belongs to the nearest-centroid type [1]. Oncotype-DX also employs a linear classifier, but it has not been described in detail.

Simplified explanation of diagnostic algorithm

When the number of diagnostic genes is n, each tumor is represented by a point in an n-dimensional space made from gene expression values. Diagnostic algorithms (linear classifier) make an $(n-1)$-dimensional plane in the n-dimensional space to separate two patient groups, e.g., high-risk and low-risk groups. As mentioned earlier, the main feature of IVDMIA is the use of a score calculated from expression values of all the diagnostic genes. I present a simplified explanation for calculation of the score using two diagnostic genes ($n = 2$).

When $n = 2$, each case is represented by a point in a two-dimensional plane made with expression of genes 1 and 2. Two coordinates correspond to expression values of genes 1 and 2. As shown in Fig. 2a, red cases (those belonging to the good prognosis group) and black cases (those belonging to the poor prognosis group) make clusters, respectively, and a border line can be drawn. It should be noted that this border line is determined by the learning set and the algorithm used.

With the coordinates in Fig. 2a, each case is represented by expression values of genes 1 and 2, e.g., (x_1, x_2). To convert these two values into a single diagnostic score (DS), the coordinates are rotated by the angle θ so that one axis (the score axis) is perpendicular to the border line (Fig. 2b). In the new coordinates, (x_1, x_2) in the old coordinates is converted to $(x_1\cos\theta - x_2\sin\theta, x_1\sin\theta + x_2\cos\theta)$. Assigning the value of the score axis at the border line is b, two groups are classified as follows.

$$a_1x_1 + a_2x_2 \geq b : \text{good prognosis group}$$
$$a_1x_1 + a_2x_2 < b : \text{poor prognosis group}$$
$$a_1 = \sin\theta, a_2 = \cos\theta$$

Thus, $a_1x_1 + a_2x_2$ acts as the score of the diagnosis.

The above example is for $n = 2$. For greater n values, DS can be simply extended as

$$\mathrm{DS} = \sum_{i=1}^{n} a_i x_i.$$

In this formula, x_i is the gene expression value for gene i, and a_i is a coefficient determined by the diagnostic algorithm with the learning data set. By defining the threshold, DS can be used to classify patients into two groups.

With oncotype-DX, a diagnostic score, named the recurrent score, is used in Paik et al. [3]. Their recurrent score is a sum of gene expression values multiplied by

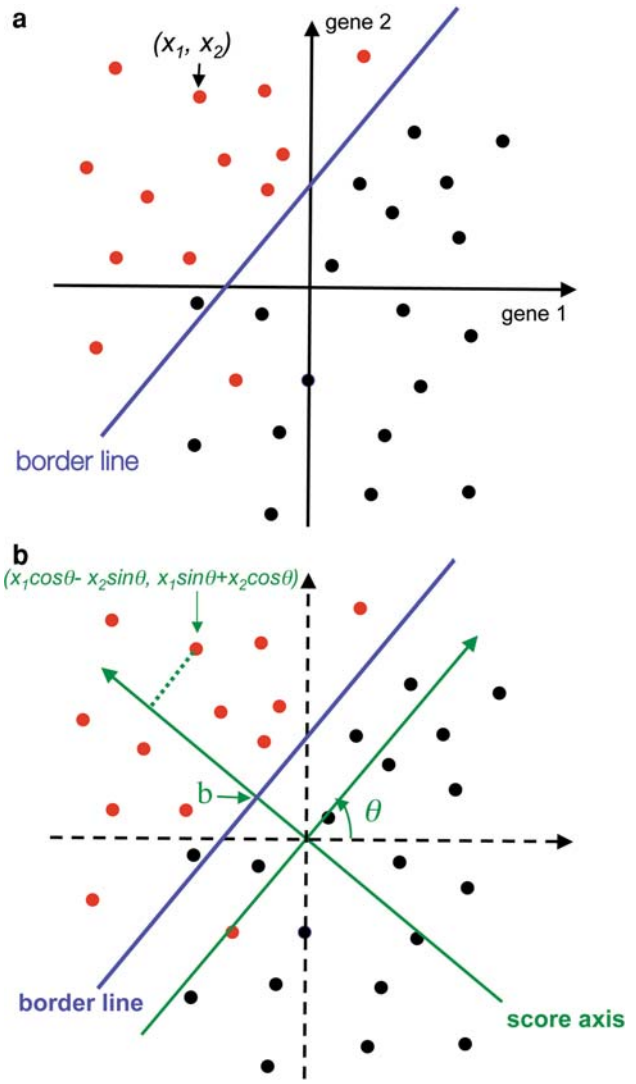

Fig. 2 Schematic representation of classification with a diagnostic algorithm. Two diagnostic genes are used. **a** Each *dot* represents a case (patient) in a two-dimensional space made from expression gene 1 and gene 2. Axes gene 1 and gene 2 indicate the expression of each gene. *Red dot*, good prognosis group; *black dot*, poor prognosis group. The borderline is constructed to separate the two groups optimally with a learning set. **b** New coordinates after axis rotation so that one axis (score axis) is perpendicular to the border line

coefficients, and conforms to the above formula. The algorithm for MammaPrint can be converted to a form of the above formula.

Additional comments

It should be noted that the above explanation is for two genes, and the process to yield a diagnostic score is somewhat different with a larger n, depending on the algorithm. However, for a linear classifier, the classification is done with an $(n-1)$-dimensional plane in the n-dimensional space made with n gene expression values. The difference is within the process of dimension reduction to yield diagnostic score.

From its definition, IVDMIA includes gene expression as well as protein expression. Large-scale identification of a protein can be achieved with mass spectrometry. This type of analysis is called proteome analysis. Proteome analysis has been mainly applied to blood samples for detection of early cancer. In spite of early studies reporting successes, e.g., a study on ovarian cancer [9], this approach is now viewed skeptically. There have been several technical advances, but more time is required to make proteome analysis plausible [10, 11]. Thus, under the current situation, IVDMIA is limited to gene expression.

MammaPrint and oncotype-DX are prognostic predictors, and will be used to provide information for decision on chemotherapy. On the other hand, there is another approach, i.e., direct prediction of effects of a particular anticancer drug. This approach has been taken with a single drug (docetaxel) [12, 13] or combined chemotherapy [14]. However, because the life of a particular regimen is short, it is not possible to recruit enough patients to establish diagnostic systems. Thus, direct prediction of anticancer drug efficacy is not a popular approach anymore.

When it first appeared, DNA microarray was expected to revolutionize medical science. The expectation was exaggerated, and now we know its limitation. However, as demonstrated by MammaPrint, it can be a powerful diagnostic tool. We are also developing IVDMIA for prognosis prediction of glioma [15, 16], which is expected to help clinical decision on temozolomide, a new alkylating agent. IVDMIA will be one of the main tools for personalized medicine.

References

1. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
2. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002;347:1999–2009.
3. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.
4. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99:147–57.
5. Kato K, Ishii S. Statistical analysis of gene expression profiles. Tanpakushitsu Kakusan Koso. 2003;48:2300–9.
6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531–7.

7. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning; data mining, inference and prediction. New York: Springer; 2001.
8. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7:673–9.
9. 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.
10. Service RF. Proteomics. Will biomarkers take off at last? Science. 2008;321:1760.
11. Service RF. Proteomics. Proteomics ponders prime time. Science. 2008;321:1758–61.
12. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge R, et al. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet. 2003;362:362–9.
13. Iwao-Koizumi K, Matoba R, Ueno N, Kim SJ, Ando A, Miyoshi Y, et al. Prediction of docetaxel response in human breast cancer by gene expression profiling. J Clin Oncol. 2005;23:422–31.
14. Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol. 2006;24:4236–44.
15. Shirahata M, Iwao-Koizumi K, Saito S, Ueno N, Oda M, Hashimoto N, et al. Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. Clin Cancer Res. 2007;13:7341–56.
16. Shirahata M, Oba S, Iwao-Koizumi K, Saito S, Ueno N, Oda M, Hashimoto N, et al. Using gene expression profiling to identify a prognostic molecular spectrum in gliomas. Cancer Sci. 2009; 100:165–72.