Prediction of cancer outcome with microarrays: a multiple random validation strategy

Lancet 2005; 365: 488–92

See Comment page 454

Stefan Michiels, Serge Koscielny, Catherine Hill

Biostatistics and Epidemiology Unit (S Michiels MSc, S Koscielny PhD, C Hill PhD), **Functional Genomics Unit** (S Michiels), **and Inserm U605** (S Koscielny), **Institut Gustave Roussy, Villejuif, France**

Correspondence to:
Dr Serge Koscielny, Biostatistics and Epidemiology Unit, Institut Gustave Roussy, 39 rue Camille Desmoulins, 94805 Villejuif, France
koscielny@igr.fr

Summary

Background **General studies of microarray gene-expression profiling have been undertaken to predict cancer outcome. Knowledge of this gene-expression profile or molecular signature should improve treatment of patients by allowing treatment to be tailored to the severity of the disease. We reanalysed data from the seven largest published studies that have attempted to predict prognosis of cancer patients on the basis of DNA microarray analysis.**

Methods **The standard strategy is to identify a molecular signature (ie, the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients. We expanded this strategy (based on unique training and validation sets) by using multiple random sets, to study the stability of the molecular signature and the proportion of misclassifications.**

Findings **The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. For all but one study, the proportion misclassified decreased as the number of patients in the training set increased. Because of inadequate validation, our chosen studies published overoptimistic results compared with those from our own analyses. Five of the seven studies did not classify patients better than chance.**

Interpretation **The prognostic value of published microarray results in cancer studies should be considered with caution. We advocate the use of validation by repeated random sampling.**

Introduction

The expression of several thousand genes can be studied simultaneously by use of DNA microarrays. These microarrays have been used in many specialties of medicine. In oncology, their use can identify genes with different expressions in tumours with different outcomes.[1–9] These gene-expression profiles or molecular signatures are expected to assist in the selection of optimum treatment strategies, by allowing therapy to be adapted to the severity of the disease.[10] Gene-expression profiling is already being used in clinical trials to define the population of patients with breast cancer who should receive chemotherapy. Such trials are being launched in Dutch academic centres and in the USA.[11]

A major challenge with DNA microarray technology is analysis of the massive data output, which needs to account for several sources of variability arising from the biological samples, hybridisation protocols, scanning, and image analysis.[12] Diverse approaches are used to classify patients on the basis of expression profiles: Fisher's linear discriminant analysis, nearest-centroid prediction rule, and support vector machine, among others.[12,13] To estimate the accuracy of a classification method, the standard strategy is via a training–validation approach, in which a training set is used to identify the molecular signature and a validation set is used to estimate the proportion of misclassifications.

Leading scientific journals require investigators of DNA microarray research to deposit their data in an appropriate international database,[14] following a set of guidelines (Minimum Information About a Microarray Experiment[15]). This approach offers an opportunity to propose alternative analyses of these data. We have taken advantage of this opportunity to analyse different datasets from published studies of gene expression as a predictor of cancer outcome. We aimed to assess the extent to which the molecular signature depends on the constitution of the training set, and to study the distribution of misclassification rates across validation sets, by applying a multiple random training-validation strategy. We explored the relation between sample size and misclassification rates by varying the sample size in the training and validation sets.

Methods

Data sources

All microarray studies of cancer prognosis published between January, 1995, and April, 2003, were reviewed in 2003 by Ntzani and Ioannidis.[1] From this review, we selected studies on survival-related outcomes (disease-free, event-free, or overall survival), which had included at least 60 patients (table). These studies used various classification methods: linear discriminant analysis, support vector machines, and prediction rules based on Cox's regression models. The sample size varied between 60 and 240 and the percentage of events between 14% and 58%.

Data were publicly available for seven studies[2–9] (webtable at http://image.thelancet.com/extras/04art5032webtable.pdf). We defined a binary clinical

Study reference	Cancer type	Clinical endpoint	Sample size	Number of events (%)	Number of channels (type)	Number of genes after filtration*
2	Non-Hodgkin lymphoma	Survival	240	138 (58%)	2 (Lymphochip)	6693
3	Acute lymphocytic leukaemia	Relapse-free survival	233	32 (14%)	1 (Affymetrix)	12 236
4	Breast cancer	5-year metastasis-free survival	97	46 (47%)	2 (Agilent)	4948
5	Lung adenocarcinoma	Survival	86	24 (28%)	1 (Affymetrix)	6532
6,7	Lung adenocarcinoma	4-year survival	62†	31 (50%)	1 (Affymetrix)	5403
8	Medulloblastoma	Survival	60	21 (35%)	1 (Affymetrix)	6778
9	Hepatocellular carcinoma	1-year recurrence-free survival	60	20 (33%)	1 (Affymetrix)	4861

*For the data of van 't Veer and colleagues,[4] the same filter was used as in the original publication. For other studies, genes with little variation in expression were excluded. †Only patients with clinical follow-up of at least 4 years after surgical resection were analysed.[7]

Table: **Description of eligible studies ordered by sample size**

outcome as described in the table. The binary endpoint was the same as in the original papers in five studies.[3,4,7–9] For the other studies,[2,5] we used the binary status of patients being dead or alive at last follow-up, instead of the time to events used by the study investigators. For all studies, we merged the training and validation sets to select training-validation sets repeatedly and randomly.

Statistical analysis

First, we eliminated genes that showed little or no variation across samples (table).[12] For every study, we divided the dataset (size N) using a resampling approach into 500 training sets (size n) with n/2 patients having each outcome, and 500 associated validation sets (size N–n). Selection of training sets including half the patients with and half without a favourable outcome maximises the power of the comparison between average gene expressions in the two groups. We identified a molecular signature for each training set and estimated the proportion of misclassifications for each associated validation set. We used different n values, from ten to a maximum value, which was chosen so that the validation set had at least one patient representing each outcome.

For a given training set, the molecular signature was defined as the 50 genes for which expression was most highly correlated with prognosis as shown by Pearson's correlation coefficient. We defined two average profiles (favourable and unfavourable) as vectors of the average expression values of these 50 signature genes in patients with favourable and unfavourable prognoses. We classified each patient in the corresponding validation set according to the correlation between expression of his or her signature genes and the two average profiles; the predicted category was that with the highest correlation. This simple method is commonly known as the nearest-centroid prediction rule.[13]

Role of the funding source

The sponsor of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

We estimated thousands of signatures (500 for every training-set size) for each of the seven microarray studies and saw that the list of 50 genes that had the highest correlations with outcome was very unstable. For instance, with data from the study by van 't Veer and colleagues[4] and a training set of the same size as in the original publication (n=78), only 14 of 70 genes from the published signature were included in more than half of our 500 signatures (figure 1). Also, ten genes not included in the published signature were selected in more than 250 of our signatures. Furthermore, 564 different genes of 4948 considered by the researchers of the original publication were included in at least one estimated signature.

Similarly, when microarray data from Iizuka and colleagues[9] and a training set of 34 patients were reanalysed, only four of 12 published signature genes were seen in more than 250 of our signatures, whereas nine not present in the published signature were also selected in more than 250 estimated signatures (figure 1). These results show how the molecular signature strongly depends on the selection of patients in the training set: we noted that every training set of patients led to a different list of genes in the signature.

Figure 2 shows the proportion (and 95% CI) of misclassifications as a function of the training-set size. With the smallest training set (ten patients), the proportion of misclassifications for the seven studies varied between 40% and 50%. For all but one study, the proportion of misclassifications decreased as the training-set size increased. This finding suggests that the proportion of misclassifications (and hence the predictive ability of the molecular signature) could be improved with large training-set sizes. The lowest proportion of misclassifications (31%) was obtained in the study of van 't Veer and colleagues[4] for a training set of 90 patients.

An upper 95% confidence limit of less than 50% for the misclassification rate suggests a significantly better predictive ability of the molecular signature than

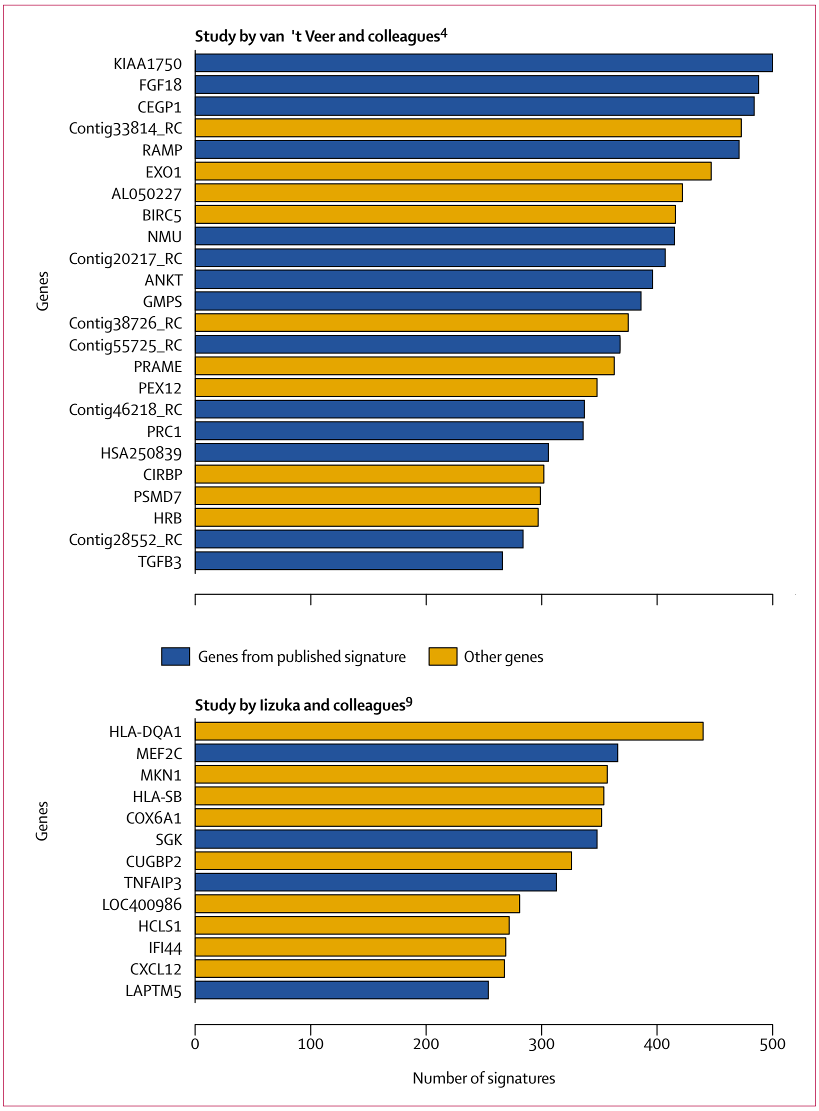

Figure 1: **Genes included in at least 250 of 500 molecular signatures for two of the studies**

expected by the play of chance. However, the 95% CI for the proportion of misclassifications fell to below 50% for some training-set sizes in only two of the studies[3,4] (figure 2). The CIs for the proportion of misclassifications were wide, emphasising the instability of estimations based on a single validation set; by definition, any individual estimate has a 95% chance of being included in the 95% CI.

Some studies published misclassification rates that were obtained by application of their classification rule to an independent validation set. These rates were taken from publications[1] and are shown in figure 2. For the studies by Rosenwald,[2] van 't Veer,[4] and their colleagues, published misclassification rates were below the lower 95% confidence limit obtained by random validation. A second validation study from the van 't Veer group reported a misclassification rate very similar to our average estimate.[16] The published misclassification rate in Beer and co-workers' study[5] was also close to our average rate. Finally, in Iizuka and colleagues' study,[9] two different classification methods were tested: the estimate from the support vector machine[12,13] was very similar to the mean classification rate obtained with our multiple random validation strategy, whereas the more data-driven score system led to an estimate below the lower 95% confidence limit.

We did a sensitivity study using other strategies to identify signature genes: selection of the 20 or 100 most discriminating genes (instead of 50) or selection of all genes with a significant correlation ($p<0\cdot01$) between expression and outcome. These three strategies yielded curves that were very similar to those in figure 2 (webfigure http://image.thelancet.com/extras/04art5032webfigure.pdf).

Discussion

We noted unstable molecular signatures and misclassification rates (with minimum rates between 31% and 49%). We used a basic algorithm to select signature genes in the training sets and an easy-to-comprehend method to classify patients in validation sets. The signature was defined by the 50 genes that were most highly correlated with the outcome in the training set. The sensitivity analyses show that our multiple random validation strategy led to results that were insensitive to changes in the number of genes selected.

We classified each patient in the validation set to a prognostic category according to the highest correlation between the expression of his or her signature genes and the favourable or unfavourable profile (defined as the average expression of signature genes in the corresponding category of the training set). This algorithm was closely similar to the method used by van 't Veer and colleagues,[4] but slightly less arbitrary: we classified patients in the validation set by the nearest-centroid prediction rule, whereas van 't Veer and co-workers classified them according to whether the correlation with the unfavourable profile was greater than 0·4. We chose a binary endpoint, favourable versus unfavourable outcome, as used in five of the studies; this endpoint ignored the timing of events. Cox's regression models could be used to take time to events into account.

In principle, there is no biological or mathematical reason why one particular classification method should be better than others for the prediction of the outcome of cancer patients by use of microarray data. Different algorithms used to classify tumours based on gene expression have been compared by Dudoit and colleagues.[17] The study included well known classification methods, such as nearest-neighbour classifiers, linear discriminant analysis, and classification trees, but also recent machine-learning

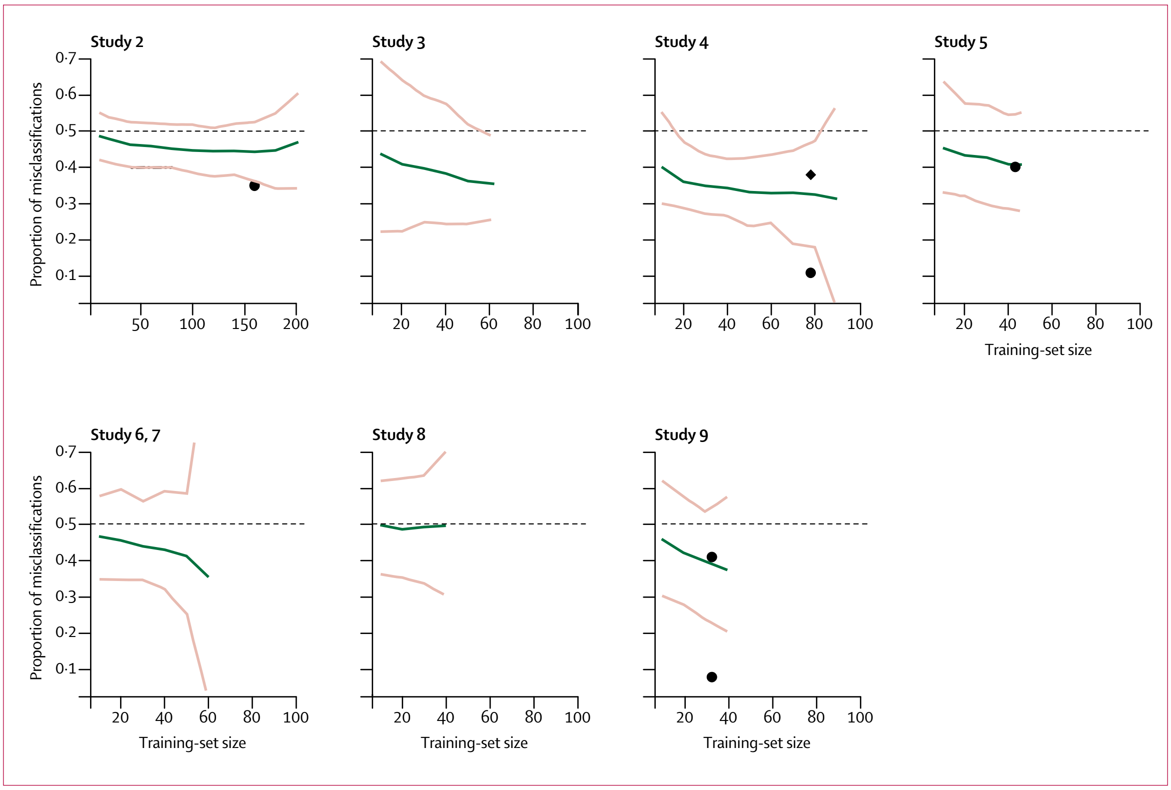

Figure 2: **Proportion of misclassifications in validation sets as a function of corresponding training-set sizes in the seven studies**[2–9]
Green lines=mean proportion of misclassifications obtained from 500 random training-validation sets as a function of the training-set size. Pale red lines=95% CIs. Dots=misclassification rates in original publications. Iizuka and colleagues[9] published two misclassification rates by two different methods on the same validation set. Diamond=second misclassification rate on a larger independent validation set[16] from the van 't Veer study.[4]

techniques such as bagging and boosting (which are supposed to improve classification by using perturbed versions of the training set).

The simplest of these methods, diagonal linear discriminant analysis and nearest-neighbour classification, predicted just as well as and even better than the complicated ones. The prediction rule used in our study, the nearest-centroid method, is very similar to diagonal linear discriminant analysis; the only difference is that our method assumes that all gene expressions have the same variability.

In conclusion, the list of genes included in a molecular signature (based on one training set and the proportion of misclassifications seen in one validation set) depends greatly on the selection of the patients in training sets. Five of the seven largest published studies addressing cancer prognosis did not classify patients better than chance. This result suggests that these publications were overoptimistic. We advocate the use of validation by repeated random sampling. Studies with larger sample sizes are needed before gene expression profiling can be used in the clinic.

Contributors

S Michiels, S Koscielny, and C Hill contributed to the conception of the study, statistical analysis of the data, and writing of the paper.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments
Part of this work was supported by the Cancéropôle Ile-de-France. We thank L Saint-Ange for editing the report.

References

1 Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 2003; **362:** 1439–44.

2 Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002; **346:** 1937–47.

3 Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002; **1:** 133–43.

4 van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415:** 530–36.

5 Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002; **8:** 816–24.

6 Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001; **98:** 13790–95.

7 Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. *Nat Genet* 2003; **33:** 49–54.

8 Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 2002; **415:** 436–42.

9 Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003; **361:** 923–29.

10 Caldas C, Aparicio SA. The molecular outlook. *Nature* 2002; **415:** 484–85.

11 Kallioniemi O. Medicine: profile of a tumour. *Nature* 2004; **428:** 379–82.

12 Miller LD, Long PM, Wong L, Mukherjee S, McShane LM, Liu ET. Optimal gene expression analysis by microarrays. *Cancer Cell* 2002; **2:** 353–61.

13 Simon R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br J Cancer* 2003; **89:** 1599–604.

14 Microarray Gene Expression Data (MGED). A guide to microarray experiments—an open letter to the scientific journals. *Lancet* 2002; **360:** 1019.

15 Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet* 2001; **29:** 365–71.

16 van de Vijver MJ, He YD, van 't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347**: 1999–2009.

17 Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002; **97:** 77–87.