available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Gene expression profiling: Does it add predictive accuracy to clinical characteristics in cancer prognosis?

Daniela Dunkler[a,d], *Stefan Michiels*[b,c,d], *Michael Schemper*[a,d,*]

[a]*Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria*
[b]*Unit of Biostatistics and Epidemiology, Institut Gustave Roussy, 39 rue Camille Desmoulins, F-94805 Villejuif Cedex, France*
[c]*Department of Biostatistics, Institut National du Cancer, France*

ARTICLE INFO

Article history:
Received 19 September 2006
Received in revised form
20 November 2006
Accepted 28 November 2006
Available online 25 January 2007

Keywords:
Cancer outcome
Explained variation
Gene expression
Predictive accuracy
Prognosis
Prognostic factors

ABSTRACT

It is widely accepted that gene expression classifiers need to be externally validated by showing that they predict the outcome well enough on other patients than those from whose data the classifier was derived. Unfortunately, the gain in predictive accuracy by the classifier as compared to established clinical prognostic factors often is not quantified. Our objective is to illustrate the application of appropriate statistical measures for this purpose. In order to compare the predictive accuracies of a model based on the clinical factors only and of a model based on the clinical factors plus the gene classifier, we compute the decrease in predictive inaccuracy and the proportion of explained variation. These measures have been obtained for three studies of published gene classifiers: for survival of lymphoma patients, for survival of breast cancer patients and for the diagnosis of lymph node metastases in head and neck cancer. For the three studies our results indicate varying and possibly small added explained variation and predictive accuracy due to gene classifiers. Therefore, the gain of future gene classifiers should routinely be demonstrated by appropriate statistical measures, such as the ones we recommend.

1. Introduction

For the last 30 years, clinical characteristics of cancer patients have been used to derive individualised predictions by means of Cox proportional hazards[1] and logistic[2] regression models. In recent years, the information derived from gene expression profiling has been used for these purposes also. The early papers promising the prediction of cancer outcome from 'gene-expression classifiers', i.e. sets of genes or signatures associated with prognosis together with classification rules,[3–6] immediately generated the impression of a major breakthrough. Later, this enthusiasm was tempered after re-evaluation of what actually had been achieved.[7–9] The merits of the pioneering studies using information from expression profiling should not be denigrated and we will likely see more successful attempts in this direction in the future. However, it has become obvious that the actual gain in predictive precision due to the use of gene classifiers derived from expression profiling needs to be carefully evaluated.

Before routine use of such potentially prognostic information can be considered, an internal validation of the gene classifier needs to be performed by means of a resampling technique such as cross-validation.[10] Then, the prognostic model based on gene expression has to be externally validated by providing evidence that the model works satisfactorily on other patients than those from whose data it was derived.[11] Third, as Kattan[12,13] notes, the predictive ability of the multivariable model that contains the marker, more specifically

* *Corresponding author:* Tel.: +431 40400 6689; fax: +431 40400 6687.
E-mail address: michael.schemper@meduniwien.ac.at (M. Schemper).
[d] These authors contributed equally to the paper.
0959-8049/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2006.11.018

the gene classifier, and other established prognostic factors should be compared to the predictive ability of the model that lacks this marker. This comparison permits an intuitive judgement of the clinical relevance of a gene-signature and is the topic of our contribution.

The objective of this paper is to show how appropriate statistical tools need to be used to answer the important question: do gene classifiers add predictive accuracy to clinical characteristics for the prognosis of cancer patients? This work was motivated by the fact that many publications in the microarray literature only show that gene classifiers are significantly associated with outcome without proper estimation of the actual gain in predictive accuracy.

Leading scientific journals require investigators of DNA microarray research to deposit their data in an appropriate international database, following a set of guidelines (minimum information about a microarray experiment).[14,15] This provides a rather unique opportunity in medical research to propose alternative analyses of the original data sets or to propose new statistical methodology. We chose three different published gene classifiers from the literature that were validated in an independent patient series and for which both clinical and genetic characteristics were available. Two of these studies[5,6] are probably the most well known examples of gene classifiers that aimed at prediction of survival of cancer patients, while the third study[16] developed a gene predictor for the detection of lymph node metastases at diagnosis. For each of the three studies, we will evaluate the predictive accuracy with and without the use of gene classifiers and thus quantify the net gain from gene classifiers for clinical practice.

2. Methods

2.1. Clinical data sets

2.1.1. Data set 1: lymphoma

Rosenwald and colleagues[6] developed a 17-gene classifier of the overall survival for patients with advanced diffuse large B cell lymphoma receiving CHOP chemotherapy. A three-level 'International Prognostic Index' (IPI) based on both clinical and pathological factors is currently used for risk stratification of patients with aggressive lymphoma (low risk: IPI 0–1, intermediate: IPI 2–3 and high: IPI 4–5). We evaluated the extent to which the continuous Rosenwald gene score adds to the IPI in the prediction of overall survival in the 73 patients of the independent validation series for which the IPI values were available.

The patient data were downloaded from http://llmpp.nih.gov/DLBCL/.

2.1.2. Data set 2: breast cancer

Van't Veer[5] developed a 70-gene classifier to predict survival in young patients with stage I or stage II breast cancer. The classifier was validated in a consecutive series of 295 patients for which the clinical data were also available.[17] We considered the endpoint of distant metastasis-free survival and used the 234 patients of the validation series that were not included in the original training series. In order to derive the set of 'clinical characteristics' of independent prognostic importance, a backward stepwise regression (threshold $p < 0.05$) was applied to the available clinical variables. We evaluated the extent to which the binary gene classifier (predicted high versus low risk) adds to the retained 'clinical characteristics' – ER-status (+/−), number of lymph nodes (0,1 to 3,4+) and histological grade (1,2,3) – in the prediction of distant metastasis-free survival. The patient data were downloaded from http://microarrays.nki.nl.

2.1.3. Data set 3: head and neck cancer

The last example is a study showing that gene expression profiling could permit early detection of lymph node metastases for primary head and neck squamous cell carcinomas.[16] This binary endpoint was determined by post-operative histological assessment. We evaluated the extent to which the 102-gene classifier adds to the pre-operative clinical assessment (N0, N+) in the prediction of the occurrence of nodal metastases in the 22-patient validation series.

The patient data were downloaded from the supplementary material in the original publication.

2.2. Statistical analysis

In this paper, we quantify predictive inaccuracy of a model by the average of the absolute difference between an observed outcome and the model prediction as developed in Schemper and Henderson[18] and Schemper.[19] For each of the three data sets predictive inaccuracy, i.e. the absolute prediction error, was determined for a model without predictors (D_0), a model using standard clinical prognostic information only (D_C), a model using the gene classifier only (D_G), and a model using both types of prognostic information (D_{CG}). In addition to the predictive inaccuracy, we also determined the proportion of variation in outcome explained by these models, or 'explained variation'.[18,19] For instance, the explained variation of the model using the gene classifier only is given by $[D_0 - D_G]/D_0$. The intent of the original development of this measure was to provide an equivalent to the R^2 in linear regression and to be able to obtain comparable values when applying models of different types to the same patient data set. The standardisation by the baseline predictive inaccuracy D_0 permits comparability of explained variation between different models and also between different data sets. Explained variation ranges between 0% and 100% (perfect prediction) and predictive inaccuracy is maximal for a model without predictors and 0 for perfect prediction. Though perfect prediction rarely will be achieved in cancer studies, in particular with survival outcomes, this theoretically achievable upper limit permits an intuitive understanding of values of explained variation obtained in practice.

Similarly, the relative gain in explained variation provided by the gene classifier when added to the multivariable model based on clinical prognostic information is calculated by $[D_C - D_{CG}]/D_0$.

For all results of predictive inaccuracy and explained variation, we estimated the standard error by bootstrap[20] using 200 re-samples each. The hypothetical effect of larger sample sizes than the actual ones on the standard error was obtained by increasing the size of bootstrap re-samples. In order to support the intuitive understanding of values of explained

variation, we demonstrate how they relate to the spread of survival functions obtained from Cox regression.[1] For this purpose the regression parameter for the gene classifier, also governing the degree of separation of corresponding survival functions, was increased iteratively until data sampled from such a (hypothetical) model produced the defined values of explained variation.

Data sets 1 and 2 were analysed by proportional hazards regression[1] as implemented in Proc PHREG of SAS/STAT[21] and data set 3 was analysed by a SAS macro for logistic regression, FL,[22,23] which is able to deal with quasi-complete separation of the data. For all models the fit was checked by means of 'candidate variables' for interactions and additionally for time-dependent effects in data sets 1 and 2. The measures of predictive inaccuracy (using an indirect formulation[19]) and explained variation were estimated using the programs SUREV and RELIMP,[24] available for free download from http://www.meduniwien.ac.at/msi/biometrie/programme/frameset_programme_en.htm.

3. Results

In this section, we directly present the results obtained by predictive accuracy calculations and do not repeat the conventional results of multivariate analyses from the original papers.

3.1. Data set 1: lymphoma

The predictive inaccuracy is 0.383 for a model without predictors for the lymphoma data, and it is reduced to 0.356 if the clinical International Prognostic Index, IPI, is used (see Table 1). The gain by adding the gene classifier to IPI is quantified by a reduction of predictive inaccuracy from 0.356 to 0.319. Thus the gene classifier is far from an optimal predictor, which would reduce the predictive inaccuracy to 0. In terms of explained variation, only 7% is attributable to IPI and an additional 10% can be explained by adding the gene classifier. Thus, for this example, the relative gain in predictive accuracy by the gene classifier is modest. In Fig. 1, the prognostic capacity of the gene classifier is illustrated by means of survival functions from proportional hazards regression, conditional on the 3 levels of the IPI score.

Here we also present conditional survival functions had the gains in explained variation due to gene classifier been 33% (i.e. a reduction of predictive inaccuracy from 0.356 to 0.238), a 'great success', or 67% (i.e. a reduction of predictive inaccuracy from 0.356 to 0.116), a likely unrealistic expectation.

Note that also the overall explained variation of 17% is in the lower range of the values typically observed with prognostic factor studies of survival (10–35%).[24] It is very common that prognostic factors explain only a small to moderate fraction of the variation in the outcomes among individuals.[18,25] They could, however, be used to identify high-risk populations in order to more efficiently conduct clinical trials.

Another important issue with studies of gene classifiers is sample size. It is obvious that the standard errors of the estimates of predictive inaccuracy and explained variation for the lymphoma study are unsatisfactorily high and that they could have been reduced by larger sample sizes. Fig. 2 shows that for this study a sample size of 400 with 260 events (proportion censored and follow-up distribution unchanged) would have cut the actual standard error in half. Gains in precision with even larger sample sizes appear minimal.

3.2. Data set 2: breast cancer

In this example, the explained variation in patient outcome is 16% for the model with the clinical characteristics and 12% for the model with gene classifier only (see Table 2).

The gain from adding the 70-gene signature to the clinical characteristics is even smaller than in the previous example, and amounts to an increase in explained variation of only 3%.

In Fig. 3 the prognostic capacity of the gene signature is illustrated by means of survival functions from proportional hazards regression: for the given sample (explained variation 12%) and for the hypothetical situation had explained variation been twice as much as that observed (24%).

Similar to the previous example, a larger sample size, about 4 times the actual one, would appear optimal in terms of standard errors of predictive inaccuracy (reduction to around 0.01) and explained variation (reduction to around 2).

3.3. Data set 3: head and neck cancer

The additional explained variation due to use of the gene classifier is 48% in this example, an encouraging result. However, from Table 3 we recognise substantial uncertainty in the estimates of predictive inaccuracy and explained variation because of the small sample size of the validation series.

For this example increases in sample sizes to 66, 330 or 880 would have resulted in much smaller expected standard errors of 7, 3 and 2, respectively, for the variation explained by

Table 1 – Explained variation and predictive inaccuracy for survival in the lymphoma data set (n = 73; 48 events)

	Predictive inaccuracy		Explained variation in %	
		Standard error		Standard error
Model without predictors	0.383	±0.01	0	
Model with International Prognostic Index, IPI	0.356	±0.02	7	±5
Model with gene classifier	0.341	±0.02	11	±6
Model with IPI and gene classifier	0.319	±0.02	17	±6
Gain by adding gene classifier to IPI	0.037		10	

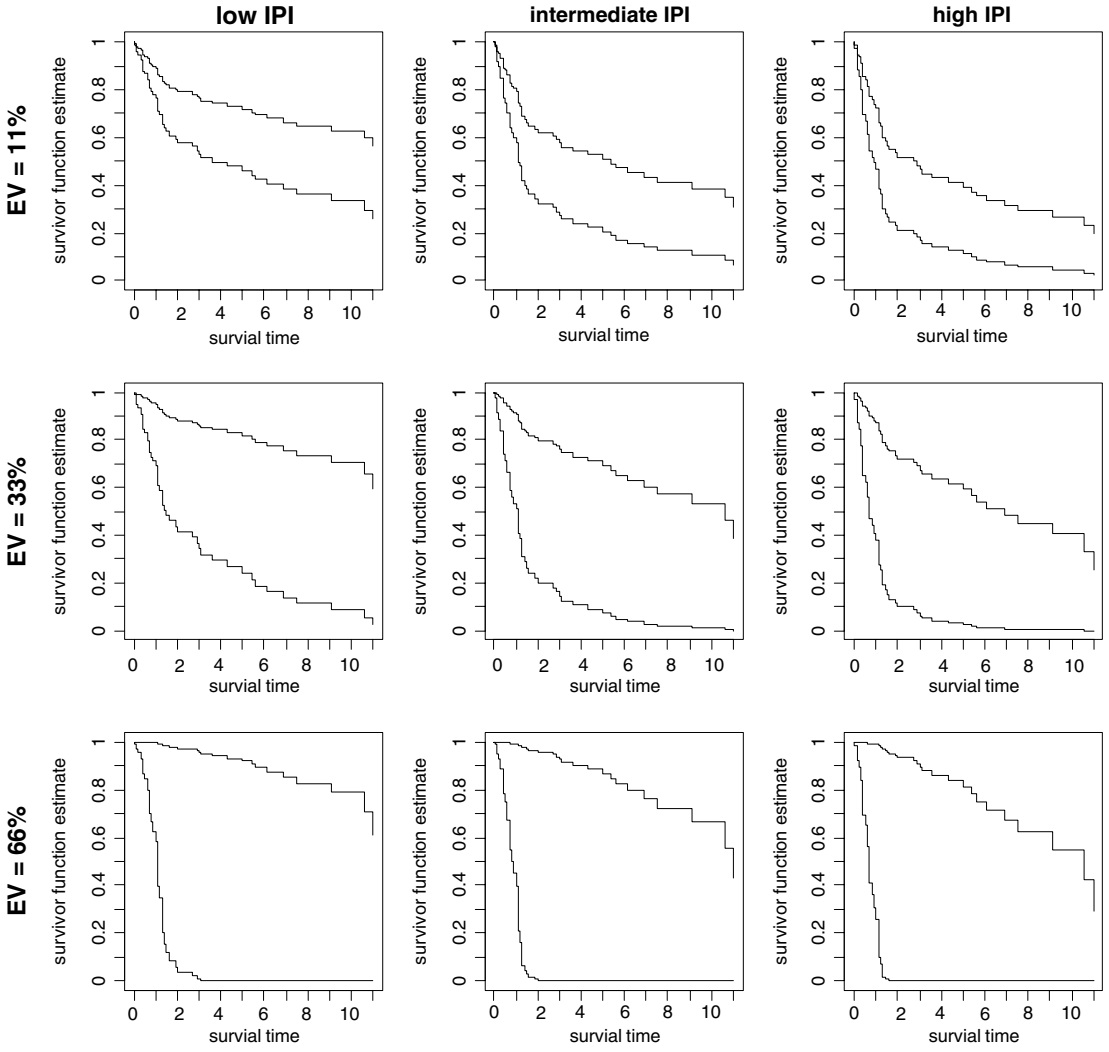

Fig. 1 – Survival functions for the upper and lower quartiles of the gene classifier, separately for the three levels of the International Prognostic Index (IPI) of the lymphoma study: upper panel: given sample; centre panel: explained variation (EV) by gene classifier set to 33%; lower panel: explained variation by gene classifier set to 67%.

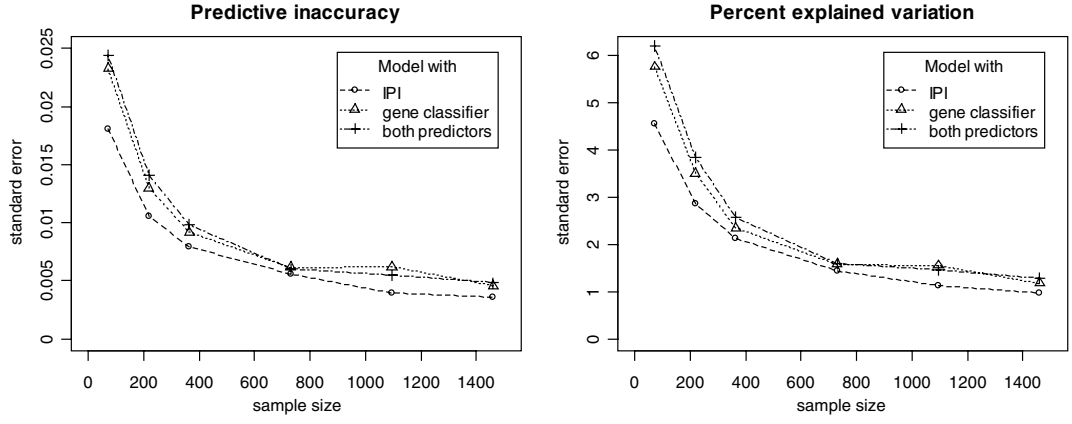

Fig. 2 – The hypothetical effect of increasing the sample size of the lymphoma study on the precision of estimates of predictive inaccuracy and explained variation as obtained by bootstrap.

Table 2 – Explained variation and predictive inaccuracy for survival in the breast cancer data set (n = 234; 55 events)

	Predictive inaccuracy		Explained variation in %	
		Standard error		Standard error
Model without predictors	0.283	±0.03	0	
Model with 'clinical characteristics'	0.238	±0.03	16	±5
Model with gene classifier	0.249	±0.02	12	±4
Model with 'clinical characteristics' and gene classifier	0.230	±0.02	19	±5
Gain by adding gene classifier to 'clinical characteristics'	0.008		3	

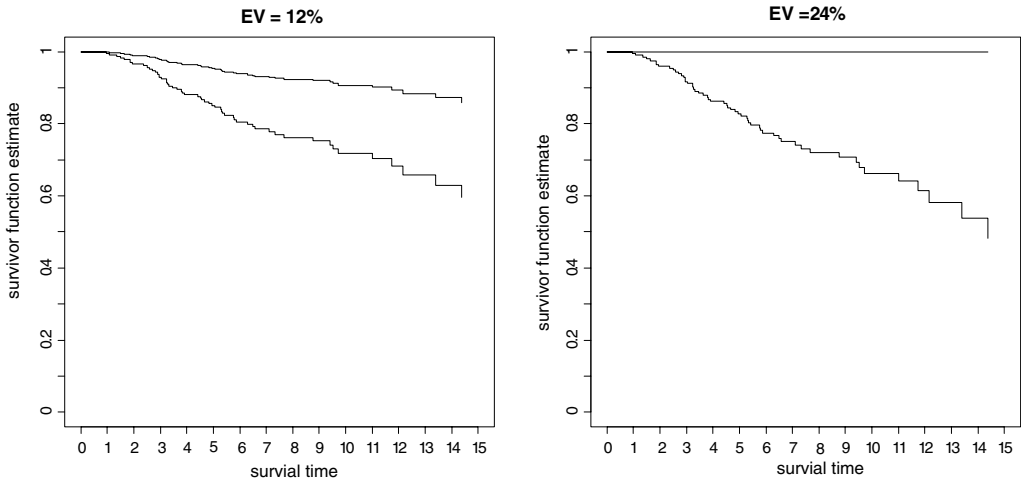

Fig. 3 – Survival functions for the binary gene classifier of the breast cancer study: left panel: given sample; right panel: explained variation (EV) by gene classifier set to 24%.

Table 3 – Explained variation and predictive inaccuracy for diagnosis of lymph node metastases in the head and neck cancer data set (n = 22; 10 metastases)

	Predictive Inaccuracy		Explained variation in %	
		Standard error		Standard error
Model without predictors	0.496	±0.03	0	
Model with clinical assessment	0.444	±0.07	10	±12
Model with gene classifier	0.260	±0.07	48	±15
Model with clinical assessment and gene classifier	0.246	±0.05	50	±10
Gain by adding gene classifier to clinical assessment	0.198		40	

clinical and genetic factors. Thus, in order to achieve reasonably precise estimates of explained variation a much larger sample of about 330 individuals would have been required.

4. Discussion

Gene expression profiling is expected to assist in the selection of optimum treatment strategies for individual patients, by allowing therapy to be adapted to the severity of the disease. In this context, it is important for medical investigators to realise that even strong and highly significant regression coefficients associated with prognostic factors of outcome may not automatically translate into sufficiently accurate prediction or close determination of individual outcome values of the patients.[19] Therefore, gains from the use of gene classifiers can only be demonstrated by the use of measures of predictive accuracy, but not by means of hazard or odds ratios, nor by their corresponding p-values. This issue is often not taken into account and even partly explains why so many identified biomarkers fail when used to predict outcomes for individual patients.[25]

In this paper, we have shown how to study gains in predictive accuracy of the prognosis of cancer patients using gene classifiers in addition to clinical characteristics. Why do we need to use a measure of explained variation in clinical studies of gene classifiers? There is an obvious danger of intuitive overestimation of predictive power when citing odds or hazard ratios. As a consequence new gene classifiers of little practical

value could be recommended for routine use. Very strong prognostic factors can have limited predictive value; for instance, for binary outcomes we learn from Pepe and colleagues[25] that a biomarker with an odds ratio of 3 is in fact a poor prediction tool and that odds ratios in the order of 30 would be desirable. But there is no intuitive interpretation of such odds ratios with respect to the gain in predictive accuracy. Alternative measures of predictive accuracy for binary and survival outcomes are the *c*-index and the area under the ROC-curve, which are identical for binary outcomes,[12,25–27] or an index of prognostic separation.[28] Since different censoring patterns of survival outcomes in the follow-up frame of interest can affect values of the *c*-index in the presence of high hazard ratios[29] but not values of the measures used by this paper, we prefer the latter ones. Neither the *c*-index nor the index of prognostic separation permits the appealing interpretation as a proportion of *variation explained* by a gene classifier. Current extensions of the ROC methodology to survival outcomes produce functions of time rather than single values.[27,30]

Occasionally, the question arises which of a few different gene classifiers developed for the same task provides the largest gain in predictive accuracy. This question can be suitably addressed by means of comparisons of predictive accuracies, and by corresponding approximate confidence intervals for their differences, for which software is readily available.[24]

For the three case studies considered here, varying and possibly small proportions of variation in the outcome could be explained by gene classifiers as compared to the established clinical prognostic factors. For example, the gene classifier for survival with breast cancer of data set 2 permitted only a reduction in inaccuracy of 0.008, from 0.238 to 0.230, amounting to a proportion of explained variation of 3%, while the gene classifier for survival with lymphoma of data set 1 permitted a not much better reduction of 0.037, from 0.356 to 0.319, related to an explained variation of 10%. In the third example the gene classifier alone explained 48% of the variability in outcome, quite a respectable result, had the sample size not been inappropriately small. Larger patient series are needed to obtain more precise estimates of gains in explained variation: for the binary example around 330 patients with balanced outcomes instead of the 22 patients would be necessary, and for the two survival examples around 260 events instead of 48 and 55 events.

To provide a rough guideline for interpreting obtained values of predictive accuracy and explained variation, we consider the effect of a gene-signature or other newly developed markers as 'weak', 'medium', 'strong' and 'very strong' if less than 20%, 20–39%, 40–59% and at least 60%, respectively, of the variation in outcome can be explained. We consider the 'medium' effect, i.e. of at least 20% explained variation, as the minimum requirement for a gain in predictive accuracy resulting in a sufficient impact on the level of an individual patient.

As mentioned in Section 1, the early papers promising prediction of cancer outcome from gene classifiers generated the impression of a major breakthrough. Our results could not confirm such a breakthrough. We do not consider such breakthroughs in predictive capacity unlikely in the near future, however, they should be critically examined by statistical tools specialised for measuring gains in predictive accuracy such as the ones we have recommended.

Contributors

All authors 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. There was no extra funding for this work and no ethics committee approval needed.

Acknowledgements

We thank C. Hill for fruitful discussion and T. Smith for correcting the manuscript.

REFERENCES

1. Hosmer DW, Lemeshow S. *Applied survival analysis*. New York: Wiley; 1999.
2. Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. New York: Wiley; 2000.
3. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7.
4. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11.
5. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.
6. 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.
7. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates an empirical assessment. *Lancet* 2003;**362**:1439–44.
8. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;**95**:14–8.
9. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;**365**:488–92.
10. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol* 2005;**23**:7332–41.
11. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.
12. Kattan MW. Evaluating a new marker's predictive contribution. *Clin Cancer Res* 2004;**10**:822–4.
13. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst* 2003;**95**:634–5.
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. Roepman P, Wessels LF, Kettelarij N, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat Genet* 2005;**37**:182–6.
17. 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.
18. Schemper M, Herderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics* 2000;**56**:249–55.
19. Schemper M. Predictive accuracy and explained variation. *Stat Med* 2003;**22**:2299–308.
20. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.
21. SAS/STAT User's Guide, Version 9. Cary (NC): SAS Institute; 2003.
22. Heinze G, Ploner M. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *CompMethProgBio* 2003;**71**:181–7.
23. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002;**21**:2409–19.
24. Heinze G, Schemper M. Comparing the importance of prognostic factors in Cox and logistic regression using SAS. *CompMethProgBio* 2003;**71**:155–63.
25. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882–90.
26. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;**247**:2543–6.
27. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;**61**:92–105.
28. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;**23**:723–48.
29. Schemper M, Stare J. Explained variation in survival analysis. *Stat Med* 1996;**15**:1999–2012.
30. Moskowitz CS, Pepe MS. Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Stat Med* 2004;**23**:1555–70.