Published in final edited form as:
Scand J Urol Nephrol Suppl. 2008 September ; (218): 185–190. doi:10.1080/03008880802283847.

Combining a molecular profile with a clinical and pathological profile: Biostatistical considerations

RICHARD J. SYLVESTER
EORTC Headquarters, Brussels, Belgium

Abstract

The use of molecular markers and gene expression profiling provides a promising approach for improving the predictive accuracy of current prognostic indices for predicting which patients with non-muscle-invasive bladder cancer will progress to muscle-invasive disease. There are many statistical pitfalls in establishing the benefit of a multigene expression classifier during its development. First, there are issues related to the identification of the individual genes and the false discovery rate, the instability of the genes identified and their combination into a classifier. Secondly, the classifier should be validated, preferably on an independent data set, to show its reproducibility. Next, it is necessary to show that adding the classifier to an existing model based on the most important clinical and pathological factors improves the predictive accuracy of the model. This cannot be determined based on the classifier's hazard ratio or *p*-value in a multivariate model, but should be assessed based on an improvement in statistics such as the area under the curve and the concordance index. Finally, nomograms are superior to stage and risk group classifications for predicting outcome, but the model predicting the outcome must be well calibrated. It is important for investigators to be aware of these pitfalls in order to develop statistically valid classifiers that will truly improve our ability to predict a patient's risk of progression.

Keywords

Area under the curve; biostatistics; molecular profile; nomograms; non-muscle-invasive bladder cancer; predictive accuracy; prognosis; progression; validation

Introduction

Non-muscle-invasive (superficial) bladder cancer (NMIBC) encompasses a heterogeneous group of patients, going from low-grade Ta tumors to high-grade T1 tumors with concomitant carcinoma in situ. Previous publications have identified a number of clinical and pathological factors related to recurrence and/or progression to muscle-invasive disease [1]. Based on an analysis of some 2500 NMIBC patients included in seven European Organization for Research and Treatment of Cancer (EORTC) trials, the EORTC developed a simple scoring system based on six routinely assessed clinical and pathological factors [prior recurrence rate, number of tumors, tumour size, stage (T category), grade and concomitant carcinoma in situ] which allows clinicians to calculate a patient's short-term and long-term risks of recurrence and progression to muscle-invasive disease [2].

Correspondence: R. J. Sylvester, EORTC Headquarters, 83 avenue E Mounier, Bte 11, 1200 Brussels, Belgium. E-mail: richard.sylvester@eortc.be.

Progression to muscle-invasive disease after intravesical treatment is associated with a greatly increased risk of death due to bladder cancer [3]. It is thus important to be able to identify accurately those patients who will progress on intravesical treatment in order to offer them an upfront cystectomy.

Although the EORTC risk tables clearly separate patients into four groups with an increasing risk of progression, their use in identifying exactly which patients will progress is somewhat limited. Among the 272 patients who progressed, 173 (63.6%) had a progression score of 7 or more, while among the 2228 patients who did not progress, 1570 (70.5%) had a progression score less than 7. Thus, using a cut-off score for progression of 7, the sensitivity (true-positive rate) was 63.6% and the false-positive rate (1 – specificity) was 29.5%. Future work needs to focus on improving the sensitivity and specificity of these tables.

Increasing attention has been paid over the past several years to the role of molecular markers and especially gene and protein expression profiling in predicting both a patient's prognosis (prognostic factor) and their response to treatment (predictive factor) [4,5]. The use of molecular markers and expression profiling provides a promising approach for improving the predictive accuracy of currently existing scoring systems such as that developed by the EORTC. The goal of this paper is to review the biostatistical aspects that must be taken into account when adding a patient's molecular profile to an existing prognostic index based on the patient's clinical and pathological characteristics. This will be illustrated for multigene expression classifiers.

Development of a multigene expression classifier

A multigene expression classifier is a function that provides the classification of a tumour based on the expression levels of the component genes [6]. Although the function could yield a continuous risk score, for validation purposes cut-off thresholds which divide patients into two or more classes, for example good risk and poor risk groups, are generally defined.

The construction of a multigene expression classifier is a multistep process: first the identification of individual genes related to the endpoint of interest and second the definition of a function used to combine the genes together into a classification rule.

Identification of individual genes

The first step in building a multigene classifier is to identify those genes that are the most differentially expressed between two or more different groups of patients, for example between those patients who progress to muscle-invasive disease and those who do not. A statistic measuring the difference in gene expression between the two groups is used to rank the genes and a cut-off is selected to produce a list of the most differentially expressed genes.

This technique suffers from a number of problems [7]. First, many false-positive genes may be identified. The false-discovery rate (FDR), the number of false positives divided the total number of positives, should be controlled; however, the false-negative rate (FNR) must also be taken into account. One proposal is to rank genes according to their *p*-value and to compare the *i*th *p*-value, p_i, to $0.05 \times i/n$, where i is the rank in the list and n is the total number of genes. Unfortunately, the FDR depends on a number of factors, many of which are not under the investigator's control, including the unknown true proportion of non-differentially expressed genes [7-9].

Secondly, gene lists may change dramatically when different patient sets are used. When patients are randomly sampled from the original patient population, a high instability in the genes included in the gene lists has been found. One technique to overcome this problem is to validate the gene list by using repeated random sampling, retaining for example only those genes that are identified in more than 50% of the resampling data sets [10].

In choosing the number of genes to be included in a classifier, a balance must be made between the risk of omitting informative genes and thus reducing classification accuracy, and including non-informative genes which may dilute the influence of the informative genes and likewise reduce prediction accuracy.

Combination of genes to develop a multigene expression classifier

The next step is to combine together the expression of the previously identified genes to develop a classifier predicting a patient's outcome. Various classification algorithms exist such as linear discriminant analysis, nearest neighbour classification, classification trees and support vector machines [6,11]. Linear discriminant analysis uses a weighted sum of the individual gene expression measurements and a threshold(s) which separates patients into one of two (or more) groups according to whether the sum is less than or greater than the threshold(s).

The method employed to define the number of genes to be included and the type of equation to be used should be defined a priori and not be based on retrospectively picking the best looking result after trying several different methods [12].

Validation of the multigene expression classifier

As the number of candidate genes that can be included in a classifier is greater than the number of patients analysed, one can always find classifiers that “accurately” classify the data on which they have been developed (training set) even if there is no relationship between the expression of any of the genes and the outcome [6].

The performance of the classifier is generally assessed based on the proportion of misclassified patients as quantified by its sensitivity and specificity. As the evaluation of the performance of the classifier on the training data set used to define it yields overly optimistic results, the classifier should be validated on data not used in developing the classifier.

External validation on an independent data set

A given classifier is optimal for the patients in the training set used to construct it. The only really correct way to validate a classifier based on the entire sample is to apply it to an external “validation” data set [13]. The conditions for the validation should be the same as those used to generate the classifier: same inclusion criteria, same endpoint, same treatment and the same definition of the classifier (list of selected genes, expression measurement, equation and cut-off value). The validation set must not include any patients used to generate the prediction rule. The performance of the classifier in the test set can be assessed in the usual way, including the calculation of the overall misclassification rate, sensitivity and specificity.

Alternatively, a split-sample validation might be carried out whereby the original sample is split into two parts, one used for developing the classifier (training set) and one used for testing or validating the classifier (test set). The adequacy of the size of the test set can be assessed by computing a confidence interval for the test set error rate.

Internal validation using resampling

Although validation on an external data set is to be preferred, when an external data set is not available, an internal validation based on resampling should be carried out. This technique involves some form of cross-validation based on repeated model development and testing on random data subsets drawn from the original sample of patients. Two popular resampling techniques using the original sample of patients are the leave-one-out cross-validation method (LOOCV) [6,14] and the leave-many-out method, where multiple training and test sets are used [10]. Each time patients are left out, the whole procedure of selecting genes and constructing the classification rule is repeated from the beginning, otherwise the percentage of misclassified patients may be greatly underestimated.

Although the classification rule to be validated may be different at each iteration of the resampling process, the model to be used for future predictions is still the one based on the entire sample. This procedure provides cross-validated estimates of the prediction error, sensitivity and specificity when using this model in future samples.

Assessment of the predictive accuracy of a gene expression classifier

Owing to selective reporting and publication bias, almost all articles on cancer prognostic markers report statistically significant results [15]. The fact that a variable is statistically significant in a multivariate model does not necessarily imply that the variable improves the model's predictive accuracy [16]. It has been shown, for example, that a marker with an odds ratio of 3 is in fact a very poor classification tool and that an odds ratio of 30 or more is desirable. More generally, however, a single measure of association such as an odds ratio does not meaningfully describe a marker's ability to classify patients [17].

Thus, the important question to answer is *not* (1) whether the gene expression classifier by itself is significantly related to the outcome, or (2) whether the gene expression classifier is statistically significant in a multivariate model which also includes the classical clinical and pathological factors, or even (3) whether the gene expression classifier is more significant than the classical clinical and pathological factors in a multivariate model, but rather: “Does adding the gene expression classifier to an existing model which is based on the most important clinical and pathological factors improve the predictive accuracy of this model?”

Neither the statistical significance (*p*-value) of the classifier in the model nor the value of its hazard ratio can be used to assess the gain in predictive accuracy. The value of the hazard ratio depends on the measurement scale and cut-off used, the other variables that are included in the model and how they are coded [18].

The following methods can be used to assess whether a gene expression classifier improves the predictive accuracy of a standard classification or scoring system when applied to an *independent* data set.

Assessment of a gene classifier within the levels of a standard risk group classification system

When a classification system exists which divides patients into different risk groups based on standard clinical and/or pathological factors, one way to assess whether the gene classifier adds predictive accuracy to this system is to examine in a separate test set the outcome for the gene classifier within the risk groups of the standard system. The statistical significance of the contribution of the new classifier within the risk groups of the standard system can be used to assess whether it adds prognostic information.

Predictive inaccuracy and proportion of explained variation

The predictive inaccuracy of a model is the average of the absolute difference between the observed outcome and outcome predicted by the model, i.e. the absolute prediction error [19-22]. The predictive inaccuracy is assessed for four models:

1. without any covariates (D_0)

2. with only the clinical and pathological factors (D_C)
3. with only the genomic classifier (D_G)
4. with both the clinical and pathological factors and the genomic classifier (D_{CG})

The proportion of variation explained by the models is then calculated. Similar to the multiple R^2 for linear regression, the explained variation for a model with only the gene classifier is (D_0–D_G)/D_0, while the relative gain in explained variation when the gene classifier is added to the model containing the clinical and pathological factors is (D_C–D_{CG})/D_0. Explained variation ranges from 0% to 100% and predictive inaccuracy is 0 for a perfect prediction. Standard errors for explained variation and predictive inaccuracy can be obtained via bootstrap resampling. The value of adding a genomic classifier to a model with clinical and pathological factors can then be determined based on the values of predictive inaccuracy and explained variation for models 2 and 4 above.

Several different modelling strategies exist, such as Cox regression, logistic regression, recursive partitioning and regression trees (CART) and artificial neural networks [23]. Considerable work has been done in defining measures of predictive accuracy and explained variation for the Cox proportional hazards regression model [20-22].

When patients die of unrelated causes before the endpoint of interest is observed, for example recurrence or progression to muscle-invasive disease, classical methods based on Kaplan–Meier estimates and Cox regression will overestimate the probability of the event. When such competing risks are present, cumulative incidence curves and special competing risk regression techniques must be used [24].

In prognostic factor studies of survival, the absolute or relative predictive accuracy and the overall explained variation are often low even when there are prognostic factors that are highly statistically significant. An explained variation of 20% has been proposed as the minimum requirement for a gain in predictive accuracy which is worthwhile on an individual patient level [19].

Area under the curve, concordance index and concordance probability estimate

The most common measures of predictive accuracy for binary and time to event endpoints are, respectively, the area under the receiver operating characteristic (ROC) curve (AUC) (sensitivity versus 1 – specificity) and the c-index (concordance index), which are identical for binary data [16,25]. The c-index, an index of predictive discrimination, provides the probability that, for two patients chosen at random, the model-predicted outcome and the observed outcome are in agreement, i.e. the patient with the worse outcome is predicted to have the worse outcome (c-index = 0.50 is agreement by chance). The c-index is calculated for the models described in the previous section to assess the amount of improvement when taking the gene classifier into account. It cannot be interpreted as a proportion of variation explained by a gene classifier.

One disadvantage of the c-index is that it tends to become more extreme as the amount of censoring increases. An alternative to the c-index which is not affected by patterns of censoring is the concordance probability estimate (CPE) based on the Cox model.

More recently, ROC methodology has been extended to time-to-event outcomes to produce time-dependent sensitivity, specificity and ROC curves along with a global concordance measure [26].

Nomograms

A nomogram provides a graphical method to predict the probability of an event, for example progression to muscle-invasive disease, based on an individual patient's prognostic score.

Nomograms are superior to staging and risk group classifications for identifying patients at high risk of an event [27]. Grouping patients assumes that all patients in a given risk group have the same prognosis and that patients on the boundaries of two adjacent risk groups have completely different prog-noses. Such groupings are likely to reduce the predictive accuracy of a prognostic model.

Before a nomogram is constructed from a model which incorporates a patient's clinical and pathological characteristics and the gene classifier, it is necessary to show that (1) the gene classifier improves the predictive accuracy of the standard model based on clinical and pathological factors, and (2) there is a good agreement between the observed event rate and the event rate predicted by the model, i.e. the predicted estimates are not biased (the model is well calibrated).

If the model with the gene classifier has not been validated on an independent data set, then an internal validation must be carried out before constructing the nomogram. To correct for overoptimism, the bootstrap resampling procedure should be used to obtain biased adjusted estimates of the concordance index and calibration curves [25].

Nomograms in bladder cancer

An overview of the use of nomograms in bladder cancer has recently been published [28]. It discusses their limitations and the various factors that should be taken into account in their design and the assessment of their performance. Nomograms have been published in both non-muscle-invasive [29] and muscle-invasive bladder cancer [30-34].

Non-muscle-invasive disease—For non-muscle-invasive disease, precystoscopy urinary levels of NMP22 improved the ability of age, gender and cytology to predict stage and grade stratified tumour recurrence [29].

Muscle-invasive disease—For muscle-invasive disease, nomograms have been developed using classical clinical and pathological factors to evaluate the precystectomy prediction of pT and pN stages at cystectomy [30] and to estimate the probabilities of recurrence [31,32] and all-cause and bladder cancer-specific survival [33] after cystectomy.

More recently, it has been shown that the assessment of multiple biomarkers (p53, pRB, p21, p27 and cyclin E1) improves the prediction of bladder cancer recurrence and death in pTa-3N0M0 patients undergoing cystectomy [34]. Internal validation was performed using the bootstrap resampling technique. This paper provides a good example of the methodology to be used when assessing whether the addition of a gene expression classifier to an existing prognostic model improves its predictive accuracy.

Conclusions

Establishing the actual benefit of multigene expression classifiers in predicting prognosis or treatment outcome presents a particular challenge at each step of the way: the identification of individual genes that are correlated with the outcome, their combination into a multigene expression classifier and its validation, and then assessing the gain in predictive accuracy when adding the classifier to a patient's clinical and pathological profile.

The statistical methodology in developing a validated classifier that is of actual benefit in day-to-day practice has received increasing attention during the past several years. The statistical techniques are only now becoming more widely known and this is an area of fertile research. In particular, two new ways of assessing improvement in model performance when adding new markers to a model, the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI), have recently been proposed as measures which expand upon the AUC [35].

It is especially important that those involved in the design of studies in this field are aware of the pitfalls and difficulties involved in order to develop statistically validated classifiers that will benefit patients and not just overly optimistic publications that will not live up to their promise.

Acknowledgements

This publication was supported by a grant from the Fonds Cancer (FOCA) from Belgium. It was also supported by grant number 5U10 CA11488-38 from the National Cancer Institute (Bethesda, Maryland, USA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Cancer Institute.

References

1. Sylvester RJ. Natural history, recurrence, and progression in superficial bladder cancer. TSW Urology 2006;1(S2):15–23.
2. Sylvester RJ, van der Meijden APM, Oosterlinck W, Witjes JA, Bouffioux C, Denis L, et al. Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. Eur Urol 2006;49:466–77. [PubMed: 16442208]
3. Schrier BP, Hollander MP, van Rhijn BWG, Kiemeney LALM, Witjes JA. Prognosis of muscle-invasive bladder cancer: difference between primary and progressive tumors and implications for therapy. Eur Urol 2004;45:292–6. [PubMed: 15036673]
4. Habuchi T, Marberger M, Droller MJ, Hemstreet GP, Grossman HB, Schalken JA, et al. Prognostic markers for bladder cancer: international consensus panel on bladder tumor markers. Urology 2005;66 (S6A):64–74. [PubMed: 16399416]
5. Bensalah K, Montorsi F, Shariat SF. Challenges of cancer biomarker profiling. Eur Urol 2007;52:1601–9. [PubMed: 17919807]
6. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23:7332–41. [PubMed: 16145063]
7. Michiels S, Koscielny S, Hill C. Interpretation of microarray data in cancer. Br J Cancer 2007;96:1155–8. [PubMed: 17342085]
8. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005;21:3017–24. [PubMed: 15840707]
9. Pawitan Y, Murthy KRK, Michiels S, Ploner A. Bias in the estimation of false discovery rate in microarray studies. Bioinformatics 2005;21:3865–72. [PubMed: 16105901]
10. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92. [PubMed: 15705458]
11. 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.
12. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumour MARKer prognostic studies (REMARK). Eur J Cancer 2005;41:1690–6. [PubMed: 16043346]
13. Altman DG, Royston P. What do we mean by validating a prognostic model? Statist Med 2000;19:453–73.

14. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics 2005;21:3301–7. [PubMed: 15905277]
15. Kyzas PA, Denaxa-Kyza D, Ioannidis JPA. Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer 2007;43:2559–79. [PubMed: 17981458]
16. Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;95:634–5. [PubMed: 12734304]
17. 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. [PubMed: 15105181]
18. Kattan MW. Evaluating a new marker's predictive contribution. Clin Cancer Res 2004;10:822–4. [PubMed: 14871956]
19. Dunkler D, Michiels S, Schemper M. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? Eur J Cancer 2007;43:745–51. [PubMed: 17257824]
20. Schemper M, Stare J. Explained variation in survival analysis. Statist Med 1996;15:1999–2012.
21. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics 2000;56:249–55. [PubMed: 10783803]
22. Schemper M. Predictive accuracy and explained variation. Statist Med 2003;22:2299–308.
23. Kattan MW. Comparison of Cox regression with other methods for determining prediction models and nomograms. J Urol 2003;170:S6–S10. [PubMed: 14610404]
24. Pintilie, M. Competing risks: a practical perspective. John Wiley and Sons; Chichester: 2006.
25. Harrell FE, Lee KL, Mark DB. Tutorial in Biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statist Med 1996;15:361–87.
26. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92–105. [PubMed: 15737082]
27. Kattan MW. Nomograms are superior to staging and risk grouping systems for identifying high-risk patients: preoperative application in prostate cancer. Curr Opin Urol 2003;13:111–6. [PubMed: 12584470]
28. Shariat SF, Margulis V, Lotan Y, Montorsi F, Karakiewicz PI. Nomograms for bladder cancer. Eur Urol 2008;54:41–53. [PubMed: 18207314]
29. Shariat SF, Zippe C, Ludecke G, Boman H, Sanchez-Carbayo M, Casella R, et al. Nomograms including nuclear matrix protein 22 for prediction of disease recurrence and progression in patients with Ta, T1 or CIS transitional cell carcinoma of the bladder. J Urol 2005;173:1518–25. [PubMed: 15821471]
30. Karakiewicz PI, Shariat SF, Palapattu GS, Perrotte P, Lotan Y, Rogers CG, et al. Precystectomy nomogram for prediction of advanced bladder cancer stage. Eur Urol 2006;50:1254–62. [PubMed: 16831511]
31. International Bladder Cancer Nomogram Consortium. Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer. J Clin Oncol 2006;24:3967–72. [PubMed: 16864855]
32. Karakiewicz PI, Shariat SF, Palapattu GS, Amiel GE, Lotan Y, Rogers CG, et al. Nomogram for predicting disease recurrence after radical cystectomy for transitional cell carcinoma of the bladder. J Urol 2006;176:1354–62. [PubMed: 16952631]
33. Shariat SF, Karakiewicz PI, Palapattu GS, Amiel GE, Lotan Y, Rogers CG, et al. Nomograms provide improved accuracy for predicting survival after radical cystectomy. Clin Cancer Res 2006;12:6663–76. [PubMed: 17121885]
34. Shariat SF, Karakiewicz PI, Ashfaq R, Lerner SP, Palapattu GS, Cote RJ, et al. Multiple biomarkers improve prediction of bladder cancer recurrence and mortality in patients undergoing cystectomy. Cancer 2008;112:315–25. [PubMed: 18008359]
35. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statist Med 2008;27:157–72.