Prognostic Models for Predicting Delayed Onset of Renal Allograft Function

M Nurminen

Citation

M Nurminen. *Prognostic Models for Predicting Delayed Onset of Renal Allograft Function*. The Internet Journal of Epidemiology. 2003 Volume 1 Number 1.

DOI: 10.5580/221d

Abstract

This review concerns methodologic issues related to prognostic modeling in medicine. To illustrate the application of the methods, prognostic models were developed and evaluated to determine important risk factors predicting delayed graft function as well as factors that prolong the delay. The data consisted of 1,215 patients transplanted with renal allografts at a single center in Finland in 1986-1995. The analysis compared commonly used risk regression and Cox regression models to newer classification tree and regression tree models. Both approaches identified the same set of most important predictors related to delayed graft function, viz. cold ischemia time, type of dialysis, and time in dialysis, but the importance ranking of other factors differed between the models. New data-driven methodologies do not necessarily offer models that are superior to the traditional ones. In complex data with many potential risk factors, prognostic tree-based models have the potential power of providing complementary information and contributing to the interpretation of prognostication.

INTRODUCTION

PROGNOSTIC RESEARCH AS A STATISTICAL CHALLENGE

Prognosis in medicine can be defined as knowledge about the probability with which the prospective course of a particular health state or event is associated with a person's profile at a particular antecedent moment in time ([1]). The profile consists of the characteristics of the person and of the disease bearing on 'background' risk, together with ones having more specifically to do with the person's propensity to experience a particular illness or other condition of ill-health.

In clinical contexts in which the illness at issue is already present, two prognostic questions can be posed (O.S. Miettinen. Personal communication): 1) Causal question: Were a certain treatment to be applied, would a certain effect (e.g. change in the course of illness) result? 2) Descriptive question: Given that the treatment is adopted, will a certain course of illness (perhaps defined by outcome only) occur? Answers to these questions are sought in terms of the proportion of instances such as the individual patient's profile. This proportion is taken as an estimate of the probability for the effect or the outcome of illness. For this probability to be meaningful to interpret, the temporal relation is either modeled explicitly as a function of time, e.g., in terms of survival time ([2]) or defined implicitly, e.g., as the duration of a surgical operation ([3]). Interpretation of these predictions assumes that, at the time of prognostication, one has the specific evidence postulated by a predictive model and the observed statistical data.

Examples of prognostic information are prediction statements about survival time (The 4-year renal graft survival and patient survival rate were 69% and 84%, respectively. ([2])) as well as about treatment response and complications of a chosen therapy (A 30% increase in systolic and/or diastolic arterial blood pressure occurred in 27% of all patients and in 67% of those who had undergone a general anesthesia. ([3])).

PROGNOSTIC MODELS IN MEDICAL RESEARCH

There are many different approaches to how prognostic models can be used, developed, and evaluated ([4]). Prognostic models are useful decision-support tools in clinical practice (e.g. in selecting appropriate treatments for individual patients), in clinical trials (e.g. in defining inclusion criteria to control for variation in prognosis), and in public heath policymaking (e.g. in deciding on the prioritizing of resource

allocation between different patient groups). There are thus two levels at which prognostic models can be useful: at population level and at individual level. It would appear that the distinction between what is achievable at the population level and at the individual level is not generally well understood. Breiman and Davidoff ([5]) put forth five criteria as a basis for the clinical appraisal of the applicability of a probabilistic prognostic model for an individual patient: 1) the comparability of the patient and the study group used to develop the model; 2) the congruence between the clinical state of interest to the patient and physician, and the model's outcome; 3) the availability of all input variables where and when the prediction is to be made; 4) the usefulness of a quantitative estimate of the predicted clinical state; and 5) the degree of uncertainty in the probability estimate. I shall not dwell on the question whether a particular antecedent was prognostic of an individual case of illness, because the concern here is with modeling on the population level.

Another distinction is that prognostic models may be developed either for practical or for scientific reasons (or both). Pragmatic studies are prompted by precise clinical aims. For example, a clinician would like to know what are the chances for his patient, a 55-year old man who recently had a successful kidney graft transplantation, of having a stable or deteriorating graft 5 years after the surgery, conditional on competing risks. In clinical epidemiology, the aim of a scientific study can be to gain knowledge about the disease process by determining which factors are associated with prognosis, or to determine which particular factor is prognostically important after controlling for other, previously identified prognostic factors. This framework is the focus of the present paper.

APPROACHES TO PROGNOSTIC MODELS

A vast body of commonly used techniques is available for building prognostic models, both deterministic and probabilistic. Examples of application range from a simple quantitative prognostic score, such as the "chronic allograft damage index" ([6]), to a multivariate logistic regression for risk prediction ([2]), as well as to Cox regression ([7]) for survival analysis, and its wide extensions to event occurrence studies in general ([8]). In occupational epidemiology, prediction of the expected number of future cases of asbestosis and lung cancer caused by asbestos exposure was done by combining a risk regression model and a simple deterministic population model ([9]). Modern regression methods involve computer search-intensive algorithms such as classification tree and regression tree procedures ([10]), neural networks ([11]), and many other machine learning (artificial intelligence) approaches ([12]). Smith ([13]) has commented that "a model is essentially a predictive machine for observable quantities". An unbiased assessment of model predictive accuracy will uncover problems that make clinical prediction models misleading or inaccurate. Harrell et al. ([14]) describe methods that are applicable to all regression models for developing clinical multivariate prognostic models and for assessing their predictive adequacy.

The notion of validating a prognostic model means that it has been shown to work adequately externally. In external validation the physician assesses whether the prognostic model works satisfactorily also for other patients than those from whose experience the model was developed ([15]). Standard statistical methods for assessing the prediction accuracy of a prognostic model include comparison of observed and predicted event rates for groups of patients, and measures which discriminate between patients who experience the outcome and those who do not. However, if external validation data are not available, there are approaches for obtaining nearly unbiased internal assessments of accuracy. The principal methods are data-splitting ([16]), cross-validation ([17]), and bootstrapping ([18]).

BACKGROUND AND OBJECTIVES

The present article deals with the development of prognostic models for use in applied medical research. The presentation that follows is an attempt to delineate appropriate approaches for prognostic modeling, and to evaluate whether the newer methods have better performance characteristics than those used commonly. A previously unanalyzed kidney transplantation data set from Finland is used for illustration. Because newer drugs have replaced the immunosupressive therapies that the studied patients received, the substantive results are no longer of clinical interest (H. Isoniemi. Personal communication), and they are presented for illustrative purposes only. Thus the outlook here is predominantly methodologic and didactic.

DATA AND METHODS

KIDNEY TRANSPLANTATION DATA

The long-term survival of kidney transplant patients and the survival of the transplanted renal allografts were previously studied by Isoniemi et al. ([2]). They obtained estimates of the survival probabilities, using survival curves, and they analyzed the influence of potential risk factors for late deterioration of renal allograft function (termed chronic

rejection or dysfunction). Despite improved graft survival, a considerable proportion of renal allografts start to function with a delay. In the transplantation data at hand, the influence of several potential risk factors related to the donor, the preservation of the graft, and the recipients were investigated. The analysis was performed in two stages. First, I considered the binary outcome of whether the onset of graft function was delayed. Second, I considered the time delay (in days) since the time of transplantation until the patient' s graft started to function. The outcome variates are referred to as 1) occurrence of delayed graft function (DGF), and 2) time to graft function (TGF).

The study population consisted of 1,215 patients with end-state kidney diseases; they were transplanted with renal cadaveric allografts at a single center between January 1986 and December 1995. All patients received initial cyclosporine therapy, continued irrespective of graft functioning. Graft failure was defined as non-life-sustaining function of the kidney requiring dialysis, retransplantation, or leading to death. Nine patients died before their graft started to function. In all, 745 patients received an organ that started to function early (i.e. within the first 24 h), 435 patients' graft functioning was delayed (i.e. more than 24 h), and in 35 cases the organ never functioned.

A number of potential risk factors related to the donor, the recipient, and transplantation procedure were evaluated: the effect of the recipient's age and gender, end-state renal disease, retransplantation, preservation or cold ischemia time (CIT), dialysis type (continuous ambulatory peritoneal (CAP) dialysis or hemodialysis), dialysis time, perfusion fluid used in the preservation of doner kidneys (University of Wisconsin (UWI) solution versus Euro Collins (EC) liquid), incompatible AB and DR matches, transplantation order, panel-reactive antibodies (PRA), the doner's age and sex, cytomegalo-virus (CMV) status, and HLA-histoincompatibility.

CMV values were not obtained in 116 consecutive cases of both donors and recipients in 1986, because the method for determining CMV was not available at that time. A complete-subject analysis, in which only subjects with all values recorded for all covariates are retained in the analysis, yields consistent results, provided that the missing data mechanism does not depend on the graft functioning or covariates ([19], [20]). However, in order not to lose information, I adopted a method in which the subjects with missing CMV values were treated as a category of their own. In the case of other covariates, there were only very few randomly missing values.

UNIVARIATE STATISTICS

Preliminary, univariate testing for the prognostic factors of the occurrence of DGF was performed by means of the Pearson Chi-squared test for proportions in the case of nominal covariates, the Cochran-Armitage test for trend in the case of ordinal covariates, and two-sample Kolmogorov-Smirnov test for numerical covariates. When TGF (number of days to graft functioning, log-transformed) was used as the outcome variate (excluding those patients with a graft that functioned early), a two-sample Kolmogorov-Smirnov test was used for grouped variates with two categories, and a one-way analysis of variance for grouped variates with more than two categories. The results were confirmed by nonparametric ('smooth') regression methods ([21], Ch. 9).

MULTIPLE REGRESSION MODELING

When the outcome was the occurrence of DGF (versus early functioning), I applied a generalized linear model ([22]) to these data. Both the logistic and exponential models for the response variate yielded essentially the same results. For dichotomous factors with a binary coding ('treatment contrasts', ([21], p.157)), the multiple logistic/exponential regression has the interpretation that its beta, β, coefficients are expressed simply in terms of odds ratio (OR) or risk ratio (RR) parameters, $\exp(\beta)$. However, because the delayed outcome was not uncommon (>30%), OR should be taken as an over-estimate of RR. (Here risk refers to the probability of DGF.) Therefore, RRs were estimated from a generalized linear model using quasi-likelihood estimation (Quasi-likelihood estimation allows one to estimate regression relationships without fully knowing the error distribution of the response variate. The theoretical likelihood function need not be exactly specified, and fewer assumptions are made in the estimation and inference) with a log link function and an estimated scaling factor (dispersion parameter) of the variance.

Before embarking on multiple regression modeling of DGF, I compared the time distributions of groups defined by the categories of single covariates (univariate analysis) using nonparametric Kaplan-Meier estimates ([23]). Pearson Chi-squared test for nominal categories, and by the test for trends for the ratio of observed to expected events for ordinal-scaled covariates ([24], Sec. 3.4) determined the significance of differences in distributions. When a covariate was numerical, its relation to the delay time was inspected by smooth regression fits to a bivariate distribution ([21], Ch. 9).

To study the joint effect of multiple factors on TGF, I used Cox regression ([7]) because the outcome is a discretized time variate (representing categories along a continuum), many patients share the same day when their graft started to function (multiple 'tied' event times). For handling ties, I used the Efron approximation ([8], p. 103). I generated simulated data, which showed that this approximation was satisfactory. In Cox regression, the exponentials of the coefficients, exp(), are interpreted as hazards or daily risks. Thus, the hazard ratio (HR) for any two subjects with covariate values x_1 and x_2 was specified as exp[(x1-x2)]. However, because the outcome actually is a favorable event (onset of graft function), I preferred to interpret the complement measure, exp[-(x1-x2)], as the HR of TGF at days after the transplantation. The assumption of proportional hazards of the Cox regression model over time was checked by graphical methods ([23]) separately for each of the potential covariates. No indication of severe non-proportionality was found.

I adopted a modeling strategy that retained several insignificant predictors. These variates were not deleted, because it would not have improved the predictive accuracy, and it would have made the confidence intervals for the estimate of or for the predicted survival probabilities with the correct coverage probabilities difficult to obtain ([14]). All computations were carried out using the S-Plus system ([25]).

PROGNOSTIC TREE-BASED MODELING

Recursive partitioning methods or tree-structured algorithms ([10], [26]) offer a non-parametric alternative to logistic and exponential regression methods. Their modular approach allows adaptation to problems such as survival analysis ([27]) and to other event analysis, with which clinicians are interested in predicting prognosis based on certain risk factors. Tree-based analysis aids in visualizing the predictive value of the significant risk factors. The results are easily conveyed to non-statisticians due to the intuitive tree structure of the predictions obtained (see Figure 3). The model is fitted using binary recursive partitioning, whereby the data are successively split along coordinate axes of the predictor variables so that, at any node, the split that maximally distinguishes the response variable in the left and the right branches is selected. The decision to branch is based on whether a so-called deviance (Deviance measures the fit of a statistical model to the data when the parameter estimation is likelihood-based, that is, it is a measure of node heterogeneity. The deviance is twice the log-likelihood of the best model minus twice the log-likelihood of the current model) exceeds a cut-off value, which is chosen to optimize the modeling. The tree construction process takes the maximum reduction in deviance over all allowed splits. Splitting continues until the nodes are homogeneous or the data are too sparse. Nodes of size 50 or larger were considered as candidates for a split, and daughter nodes must exceed size 10 for a split to be allowed. Terminal nodes are called 'leaves', while the initial node is called the 'root'. The classification tree is essentially a set of rules represented by decisional nodes that are assigned to a class.

It is also possible to have a regression tree in which each leaf gives a predicted value for the class. The tree-construction method used in the analysis can use both categorical variates and binary splits on continuous variates. No distributional assumption is required for these variates. However, the tree structure relies on dichotomization.

The tree-construction process has to be seen as a hierarchical refinement of probability models. At each node of a classification tree there is a probability distribution over the classes. The probabilities available for each terminal node remain dependent on the structure of the tree (viz. its depth). It follows that the interpretation of this probability may not be exactly the same as the one provided by the logistic model or exponential risk model. The method, as implemented in S-Plus ([26], Vol. 1, Ch. 12), allows cutting trees down in size ('pruning') using various information criteria to avoid overfitting. Alternatively, one can use cross-validation on a separate data set to choose the degree of pruning.

RESULTS

In all, 61% of the patients experienced early graft functioning, 36% had DGF, 2% of the grafts never functioned, and 8% of the patients died before their grafts started to function. Acute rejection occurred in 25% of the recipients. Currently, acute rejection is curable in most cases, and it caused death in only 3% of the cases. The median day of onset of graft functioning in the DGF group was the 10th day.

UNIVARIATE ANALYSIS

The statistical analysis of the transplantation data was performed in two stages, beginning with univariate analyses of each potential risk factor. This first step resulted in the identification of a set of significant covariates for the occurrence of DGF: recipient's sex; renal disease; order of transplantation; CIT; type of dialysis; time in dialysis; PRA

last value; PRA highest value; donor's age; AB mismatches.

The covariate set identified in the univariate analysis of TGF was: order of transplantation; time in dialysis; perfusion liquid; PRA last value; PRA highest value; donor age; AB mismatches; DR mismatches.

RISK REGRESSION ANALYSIS

The fit of the risk regression model with the whole (non-parsimonious) set of covariates was very good (residual deviance 759 on 1183 degrees of freedom). There were several significant risk factors in the occurrence of DGF (Table 1).

Table 1

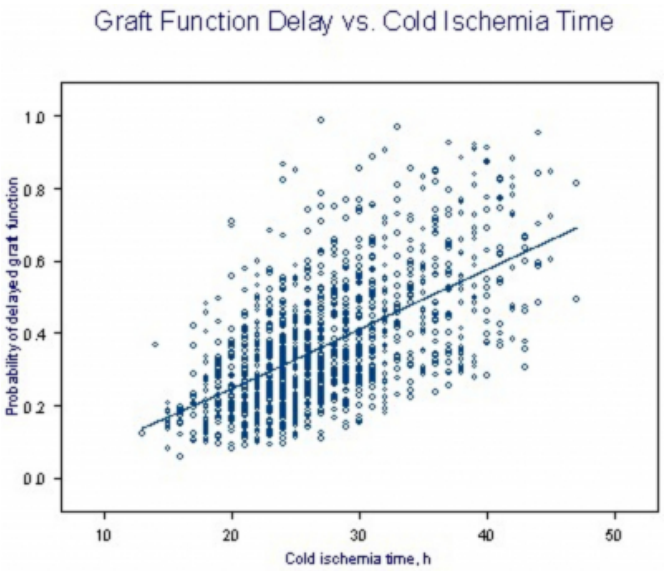
Risk ratio estimates for the occurrence of graft function delay.

Predictor variate	Risk Ratio = exp(β)	95% lower confidence limit	95% upper confidence limit	Two-sided P value
Recipient's sex:				
man vs. woman	1.21	0.81	1.80	0.35
Recipient's age:	1.006	0.999	1.013	0.07
Diagnosis of a renal disease:*				
1. Diabetes	1.00	..	..	..
2. Glomerulonephritis	0.93	0.73	1.19	0.58
3. Pyelonephritis	1.01	0.59	1.75	0.96
4. Polycystic kidneys	1.05	0.76	1.45	0.77
5. Amyloidosis	1.43	0.80	2.58	0.23
6. Other	1.21	0.91	1.61	0.19
Order of transplantation:†				
2nd	1.08	0.86	1.35	0.52
3rd	1.61	1.09	2.38	0.017
4th	1.10	0.53	2.26	0.80
Cold ischemic time, hours	1.048	1.037	1.059	<0.001
Dialysis time, months	1.005	1.001	1.009	0.001
Dialysis type:				
Hemodialysis vs. CAP dialysis	1.57	1.33	1.85	<0.001
Perfusion liquid:				
University of Wisconsin vs. Euro-Collins	0.87	0.68	1.11	0.26
PRA, highest value	1.003	0.9998	1.006	0.069
Donor's sex: male vs. female	0.98	0.84	1.14	0.79
Donor's age, polynomial:				
linear term	**	**	**	0.013
quadratic term	**	**	**	0.34
cubic term	**	**	**	0.011
Cytomegalo-virus infection:‡				
Recipient -, Donor +	1.29	0.82	2.03	0.26
Recipient +, Donor -	1.31	0.87	1.98	0.20
Recipient -, Donor –	1.26	0.81	1.96	0.31
R & D missing data	1.18	0.74	1.85	0.49
HLA-histoincompatibility:§				
AB mismatches 1-3	0.84	0.68	1.04	0.53
DR mismatches 1-2	1.06	0.88	1.28	0.10
Interaction:				
Recipient's sex × Diagnosis 2	**	**	**	0.11
Recipient's sex × Diagnosis 3	**	**	**	0.70
Recipient's sex × Diagnosis 4	**	**	**	0.03
Recipient's sex × Diagnosis 5	**	**	**	0.12
Recipient's sex × Diagnosis 6	**	**	**	0.01
Interaction: Recipient's sex × DR mismatches	**	**	**	0.026

* Relative to diabetes; † Relative to 1st transplantation; ‡ Relative to (Recipient + & Donor +);
§ Relative to no mismatches; ** Category not applicable

Figure 1

Probability of delayed graft function in relation to cold ischemia time with a smooth regression line.



The most important findings were: the risk of DGF increased linearly with increasing CIT (Figure 1); the risk was higher among hemodialysis patients than among patients who underwent CAP dialysis; and longer time in dialysis carried an increased risk.

The donor's age depicted a curvilinear relation to the estimated probability of function delay (represented in the model by a data-dependent, third-degree polynomial transformation). In the youngest age group (below 25 years) there was no increased risk. In the age span of 25-to-50 years, the risk of DGF grew with advancing age. In the oldest age group (above 50 years), the increase in risk tended to level off. However, the risk estimates were imprecise at both ends of the age distribution due to the relatively few numbers, so that the apparent nonlinearity could be a statistical artifact.

Although the number of AB mismatches was related to DGF in the univariate analysis, the relation ran counter to the hypothesis: patients with one or more AB mismatches experienced less frequently the occurrence of DGF than did patients with no mismatches. However, in the multivariate analysis the 'AB-mismatches' variate was no longer significant, but the term was nevertheless retained in the model. On the other hand, the relation of the number of DR mismatches to the short-term outcome was in line with the hypothesis. A patient who had one or two DR mismatches incurred (with 95% probability) a raised risk of DGF relative to a patient with no DR mismatches. Moreover, the effect of

DR mismatches on the risk of delayed functioning was different for men and women. Thus a product term 'Recipient sex DR mismatches' was added to the model with 'DR mismatches' represented as a binary variate (no mismatches vs. 1-2 mismatches). A female patient who had 1 or 2 mismatches carried a 1.6-fold risk relative to that of a female patient with no mismatches. For a male patient, DR mismatches did not increase the risk of graft function delay. This pattern persisted in the data when they were divided randomly into two parts as a cross-validation check.

The recipient's sex also modified the risk of DGF in the diagnostic subgroups. In patients with diabetes, the risk was 36% for both men and women. The risk in patients with polycystic kidneys, compared to that of the patients with diabetes, was higher among men and lower among women; that is, there was a significant risk difference between men and women in this diagnostic category.

When modeled jointly, not all the significant univariate factors remained significant. In prognostic modeling, however, full model fits (i.e. leaving all hypothesized variates in the model regardless of their P value) are often more adequate than fits after screening predictors for significance. For example, the PRA variate was significant in the univariate analysis, but it did not appear to be an important factor in the multivariate prognostic model. This may be because PRA was strongly associated with the order of transplantation: of the subjects receiving their 1st, 2nd, 3rd, or 4th renal allograft, 6%, 39%, 44%, and 75 %, respectively, belonged to the categrory of the highest PRA 50%. Nevertheless, I chose to keep PRA in the model for comparative purposes.

COX REGRESSION ANALYSIS

The fit of the hazard regression model yielded several significant prognostic factors for TGF (Table 2).

Table 2

Hazard ratio estimates for delay in the time-to-initiation of graft functioning.

Predictor variate	Hazard ratio = exp(-β)	95% lower confidence limit	95% upper confidence limit	Two-sided P value
Recipient's sex:				
male vs. female	1.11	0.89	1.39	0.34
Recipient's age:*				
25-34 years	1.28	0.80	2.04	0.30
35-44 years	1.37	0.87	2.15	0.18
45-54 years	1.43	0.90	2.27	0.13
55-64 years	1.62	1.01	2.59	0.045
65-74 years	0.75	0.34	1.66	0.48
Renal disease:†				
Diagnosis 2	0.79	0.59	1.06	0.12
Diagnosis 3	0.65	0.42	0.99	0.043
Diagnosis 4	0.77	0.52	1.13	0.18
Diagnosis 5	0.97	0.51	1.86	0.94
Diagnosis 6	1.13	0.80	1.60	0.49
Order of transplantation:‡				
2nd	0.91	0.67	1.25	0.57
3rd	0.99	0.54	1.82	0.98
4th	1.35	0.52	3.50	0.54
Cold ischemia time, hours	1.012	0.997	1.027	0.13
Dialysis type:				
Hemodialysis vs. CAP dialysis	1.08	0.86	1.36	0.52
Time in dialysis, months	1.012	1.006	1.018	<0.001
Perfusion liquid:				
University of Wisconsin vs. Euro-Collins	1.56	1.09	2.24	0.016
Panel-reactive antibodies, highest value	1.002	0.998	1.007	0.25
Donor's sex: male vs. female	0.97	0.78	1.20	0.79
Donor's age:§				
1-9 years	2.49	1.22	5.09	0.012
10-19 years	1.00	**	**	**
20-29 years	0.89	0.58	1.36	0.58
30-39 years	1.01	0.69	1.49	0.94
40-49 years	1.16	0.80	1.67	0.44
50-59 years	1.34	0.90	2.01	0.15
60-69 years	1.38	0.56	3.36	0.48
Cytomegalo-virus infection: ¦				
Recipient - , Donor +	1.05	0.55	2.01	0.87
Recipient + , Donor -	1.22	0.67	2.23	0.51
Recipient - , Donor –	1.18	0.63	2.20	0.62
R & D missing data	1.42	0.74	2.75	0.29
HLA histoincompatibility:¶				
AB mismatches 1	1.12	0.81	1.53	0.50
AB mismatches 2	1.28	0.93	1.77	0.13
AB mismatches 3	2.32	1.02	5.26	0.043
DR mismatches 1	1.13	0.89	1.42	0.31
DR mismatches 2	1.47	1.00	2.15	0.047

* Relative to 15-24 years; † Relative to diabetes;
‡ Relative to 1st transplantation;
*§ Relative to 10-19 years;
¦ Relative to (Recipient + & Donor +);
Relative to no mismatches; ** Category not applicable;

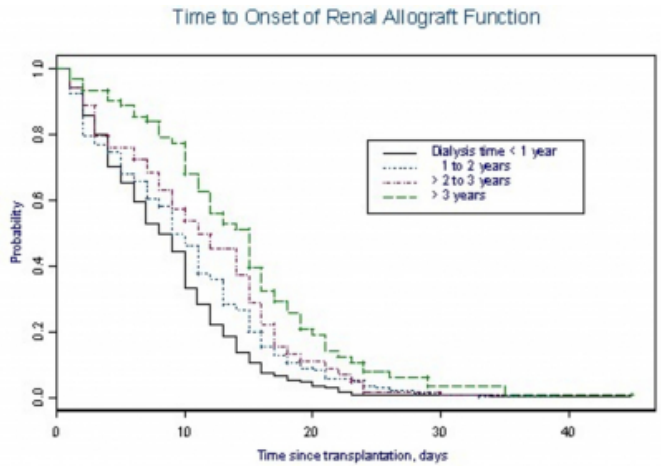
The most important finding was that longer time in dialysis incurred a higher risk. The effect of being treated during a period of 1- <2, 2- <3, and over 3 years in dialysis relative to less than 1 year duration of treatment was to increase the HR to1.3, 1.6, and 2.3, respectively. The use of the University of Wisconsin perfusion liquid instead of the Euro-Collins liquid entailed a significant increase (56%) in the hazard. When the number of DR mismatches was 2, the hazard grew by 47% relative to having none. Similarly, when the number of AB mismatches was 3, the hazard increased by 32%.

As an illustration, Figure 2 depicts the success curve for graft functioning predicted by the final Cox model: the

proportion of recipients with a non-functioning graft are plotted as a function of time following transplantation, separately for patients who underwent different durations of dialysis therapy. With increasing duration, the probability of the onset of graft function grew smaller. This trend was consistent over the entire time range.

Figure 2

Probability of time to onset of graft function for subgroups by duration of dialysis.

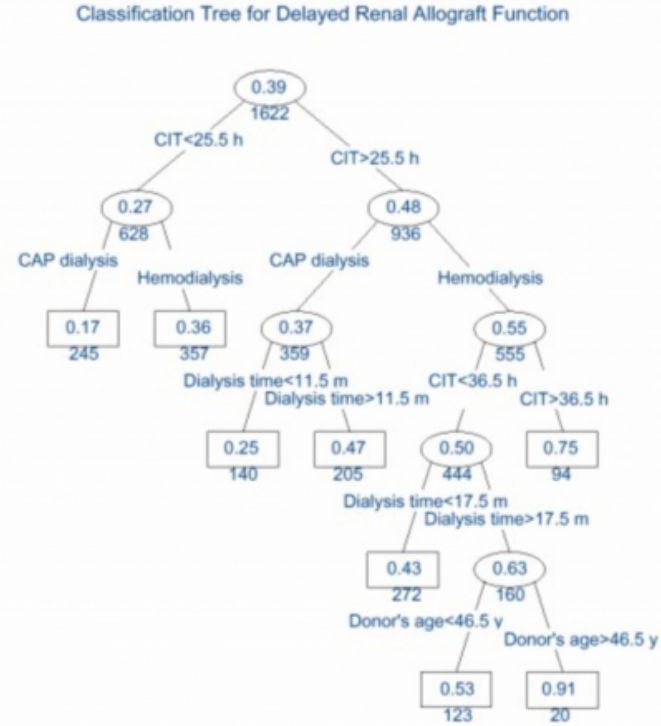


TREE-BASED ANALYSES

Figure 3 depicts a classification tree model with predicted probabilities of DGF. An interpretation of the tree is as follows. The number in the node is the model-predicted probability for DGF. At the root of the tree the predicted probability equals the observed proportion of DGF. In these data, CIT was the most important predictor. This influence, however, seemed somewhat ill-captured by the repetitive binary structure of the tree, as two splits are necessary to depict the relation. The same situation applied to the duration of dialysis treatment. This pattern may partly compensate for the binary categorization of the continuous variates, which have the potential of carrying more information.

Figure 3

A classification tree-based model for the probability of delayed renal allograft function. The edges connecting uniformly spaced interior nodes (ellipses) are labeled by left and right splits. The number in each node is the prediction for the patients; those underneath the nodes indicate the deviance contributions.



Abbreviations: CIT = cold ischemia time, CAP = continuous ambulatory peritoneal, y = year, m =month, h = hour.

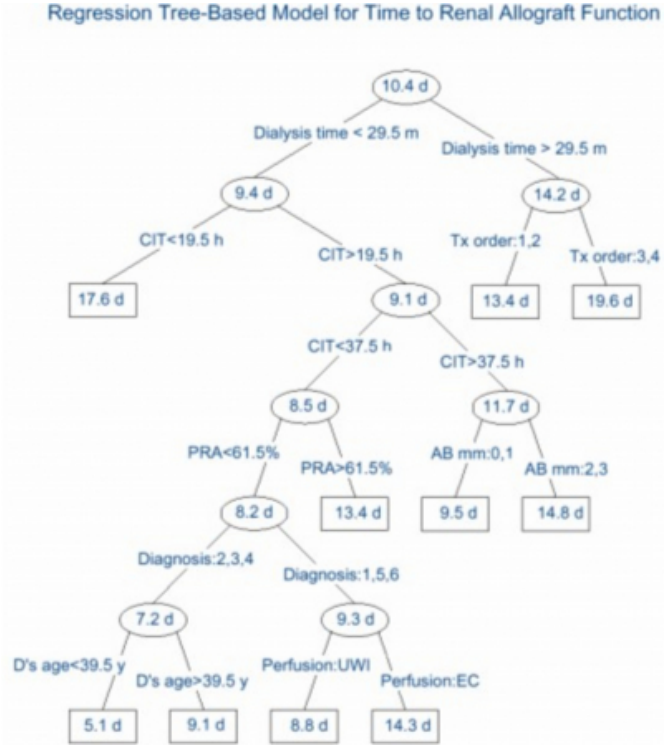
The deviance contributions are given below the nodes. At the root the deviance was 1622. The first split reduced the deviance by 1622 - 628 - 936 = 58. With the addition of nodes, the deviance was further reduced, but it was not obvious to choose the optimal degree of pruning. Using cross-validation and plotting the deviance versus the size of the tree showed that the deviance was close to the minimum when the size was 8. An increase in the size to 10 did not improve the misclassification error rate. Therefore I chose the more parsimonious model.

There was a good spread of the predicted probabilities in the 8 terminal nodes. The lowest probability of DGF, 14%, pertained to a patient (classified into the most left-side group) whose graft's CIT was less than 25.5 h, and who had had CAP dialysis for less than 29 months. On the other hand, the highest probability, 91%, was associated with a patient (classified into the most right-side group) whose graft's CIT was more than 25.5 h, who had had hemodialysis for more than 17.5 months, and whose doner was older than 46.5

years. Finally, there were indications of the presence of interactions, since the doner's age and recipient's age factors appeared only in either one of the branches of the tree. The estimated misclassification error of the predictions was 33% (= 397/1215).

Figure 4

A regression tree-based model for the time to renal allograft function.



The edges connecting uniformly spaced interior nodes (ellipses) are labeled by left and right splits. The number in each node is the predicted time in days for the patients. Abbreviations: CIT = cold ischemia time, Tx = transplantation, AB mm = AB mismatches, PRA = panel-reactive antibodies, D. age = doner's age, UWI = University of Wisconsin liquid, EC = Euro Collins liquid, y = years, m =months, d = days, h = hours.

Figure 4 displays a regression tree for TGF. The number at the root node, 10 days, is the observed average number of days from transplantation to the onset of graft functioning. The data were a subset of the 435 patients with DGF. The cut-off value for the split intercepts the line connecting adjacent nodes. The numbers in the nodes are the model predicted TGDs. The 10 terminal nodes are represented by the rectangles. The size of the tree was chosen on the basis of cross-validation. The most important predictor variates were: duration of dialysis therapy, CIT and order of transplantation. The next important predictors were the percentage of PRA and the number of AB mismatches followed by the diagnosed renal disease. Additional predictors were doner's age and the perfusion liquid used. The TGF varied in the classes from 5 to 20 days.

COMPARATIVE ANALYSIS

It is of interest to compare the results obtained by the different methods used. When the outcome at issue was DGF, both the risk regression model (as determined by the RR and P value) and the classification tree-based model (as determined by the order of inclusion into the tree) identified the same four most important predictors (viz. CIT, type of dialysis, time in dialysis, and doner's age). When the outcome variate was TDF, the hazard regression model and the regression tree-based model involved basically the same set of predictors. The most important risk factor in both analyses was the time in dialysis, whereas the importance of the other factors' rankings differed between the models. In addition, the tree structure included two factors (order of transplantation and PRA) which were not significant in the Cox survival analysis.

DISCUSSION

MODEL SELECTION STRATEGIES

Generally speaking, every model is plausible as long as it is not falsified. However, the falsification or verification of models is a challenging task in scientific research. This challenge is met both in the context of statistical models and models based on artificial intelligence approaches. To provide valid models that fulfil the task of modeling reality, it is important to consider the model building process in depth, as well as to check and validate fitted models by means of model diagnostics, whether using analytic diagnostic tools or with the aid of graphical techniques. In essence, the question is: Are we modeling the right patterns of the data with our model, and how can we validate and evaluate the model adequacy?

The process of model selection may be viewed as a pattern-recognition process ([28]). According to Greenland ([29]), to assert that a model fits the data well using some appropriate criterion in the ordinary sense means that the data appear consistent with the pattern predicted by the model. Conversely, to assert that the model fit is poor means that the data appear to deviate appreciably from the model pattern. Model correctness cannot be inferred from the fact that the fit of the model to the data is good, since there are alternative models that may also provide a good fit. As pointed out by Greenland ([29]), a good fit is, however, a

necessary condition for inference.

In the context of prognostic modeling, where the objective is interpretation, given specific states of knowledge, the function of 'automatic' procedures for the selection of variates into a model (such as stepwise forward selection) seems limited. In principle, all simple models adequately fitting the original data set (statistical validation) should be listed, and any choice between them should be made on how successfully a model performs on a new data set (clinical validation). Constructing trees may be seen as a type of variate selection procedure. Issues of interaction between variates are handled automatically. The questions are reduced to which variates to divide on, and how to achieve the split. The justification for the tree-based methodology is to view the tree as providing a probability model. The decision on a split is made based on a change in a deviance measure for the tree under a likelihood function that is conditioned on a fixed set of observed random variates. Note that once fixed by observation, a random variate is in no sense variable, so it might better be called a prognostic indicator. The unknown prediction probabilities are estimated from the proportions in the split node. The tree construction process chooses the split according to the maximum reduction in the deviance measure ([26]). When the objective is prediction, classification error-rate is the appropriate criterion for judging any particular model selection procedure.

As pointed out by Dannegger ([30]), the hierarchical ordering of the included factors lends itself to evaluation of the factor's importance. The general rule is that the closer to the root node a factor appears, the more important is its effect on the outcome. Moreover, the dichotomous representation of factors allows convenient characterizations of individuals falling below or above a certain cutpoint. On the other hand, the relation between a factor and the outcome may not be natural in the sense that the cutpoint divides the group essentially into two homogeneous subgroups regarding the factor's influence on the outcome. In these circumstances, deciding on the existence of a suitable cutpoint is problematic. Recursive partition algorithms are data-driven procedures and as such they behave, by and large, as a black-box system concerning the choice of an adequate cutpoint. Fortunately there are useful analytical and graphical diagnostic tools available for assessing the stability of the constructed tree ([30]).

PROBLEM OF OVERFITTING PREDICTIVE MODELS

There are deficiencies in the standard modeling methods. It is well known that analyses that are not prespecified but are data-dependent are liable to lead to overoptimistic conclusions ([15]). Many applications involve a large number of variates to be modeled using a relatively small patient sample. In the case of very expressive models such as regression trees, there is the danger that the models come up with chance idiosyncrasies of the study data, which are not 'true' in general. It is also important that the minimum size of the terminal nodes is sufficiently large. Problems of overfitting and of identifying important covariates are exacerbated in prognostic modeling, because the accuracy of a model is more a function of the number of events than of the sample size ([32]). For binary outcomes, Harrell et al. ([14]) suggested that in order to have predictive discrimination that validates on a new sample, the number of patients in the less frequent outcome category should exceed 10. Feinstein ([33]) suggested that the minimum number of patients should rather be more than 20. Otherwise, a data reduction technique such as deriving clinical summary indexes or variable clustering should be used for improved prognostic modeling ([25]).

An alternative approach to deal with the problem of a large number of variates is a shrinkage method that re-calibrates the model that results from overfitting ([34]). Experience indicates that a larger model is more likely to give overoptimistic prediction when extensive variate selection has been done. For example, in a study on kidney graft survival ([35]), a separate parameter was fitted for each of the 52 transplantation centers that supplied data. The researchers used empirical Bayes methods to obtain shrank estimates of the regression coefficients for the centers, which resulted in improved prediction performance.

TREE-STRUCTURED CLASSIFICATION

A tree-based prediction process produces intuitive, suggestive results. Although the risk and hazard regression models were used to predict the probabilities of DGF and TGF outcomes (Figures 1 and 2), whose distribution could be divided, say, into risk octiles or risk deciles, the tree structure defined the risk groups by their own set of prognostic factors. In addition to stratifying a study population into subgroups with distinctly different risk expectations, a tree method also permits simple and intuitive identification of potential prognostic factors and their possible interactions.

Compared to linear models, tree-based models have the following advantages: they are more apt at capturing nonadditive behavior; allow more general (i.e. other than a particular multiplicative form) interactions between predictors; invariant to monotone re-expressions of predictors; and easier to interpret when the predictors are a mix of numeric variates and indicators. Moreover, the graphical representation of the predictor set is close to medical reasoning and can be a useful tool when discussing prediction results with clinicians. Despite these appealing features, only few prognostic models have as of yet accrued adequate evidence of accuracy, generality, and effectiveness ([31]). Given the complexity of the methods used to develop and evaluate prognostic models, their usefulness in supporting clinical decision making and credibility for prognosticating illness outcomes should be assessed in collaboration with physicians. This is because for finding models that predict well, there is no substitute for understanding the substantive nature of the relations between the outcome variate and the predictor variates.

As an assessment of the stability of the tree-structured classifier, the data used for illustration was randomly split into two parts. Randomness guarantees unbiased estimation of the classification error rates. One sample was used for estimating the classification tree model. The model was then applied to the other sample for predicting the probabilities of DGF, which were dichotomized (coded as 0 = below 0.5 and 1 = above 0.5). A cross-tabulation of the model-predicted probabilities with the observed data (coded as 0 = no DGF and 1 = DGF) yielded a misclassification error rate of 35%. When the roles of the samples were exchanged, the rate became 36%. Comparison of these error rates with the apparent error rate (i.e. 33%) for the total data set shows that the latter is only slightly underestimated. This points to a stable probability structure of the tree. In other words, there is virtually no indication that the model with 10 terminal nodes would be more complex than what is necessary to describe the data.

PERFORMANCE OF MODERN VERSUS TRADITIONAL METHODS

Machine learning (i.e.modeling of data with the aid of few or no rules) is being increasingly applied for risk prediction in medicine ([36], [37]). The machine learning approach is especially suitable for discovering patterns in vast sets of biomedical data governed by complex rules. Therefore, it would be valuable to learn whether these modern methods do, indeed, have better performance characteristics than those used commonly. Lapuerta et al. ([38]) compared the accuracy of a neural network and a Cox regression model for predicting the risk of coronary heart disease, and found that the latter showed a higher classification error rate. Knuiman et al. ([39]) found that a decision tree and a logistic regression model had a similar discriminate ability to predict coronary mortality. Duh et al. ([40]) speculated that complicated classification techniques would prove useful for solving epidemiologic problems that require pattern recognition, and that they would be unfavorable for problems that involve distinct effects of distinguishable predictors. In another study, Duh et al. ([41]) suggested that more generalizable modeling techniques for neural networks might be needed before they are feasible for medical research. On the other hand, Ioannidis et al. ([42]) concluded that artificial intelligence methods might complement linear statistical methods. Ennis et al. ([43]) compared the performance of some recently developed statistical learning method for the prediction of mortality from acute myocardial infarction in a very large real database involving 41,021 patients admitted to 1,081 hospitals in 15 countries. Specifically, the evaluated methods were: neural networks, classification trees, generalized additive models, and multivariate adaptive regression splines. They found that none of the newer methods could outperform the simple logistic regression model previously developed for this problem, and consequently concluded that adaptive non-linear algorithms might have limited applicability in clinical settings. In a critical commentary, Schwartzer and Vach ([44]) concluded that there is no evidence so far that the application of neural networks represents real progress in the field of diagnosis and prognosis in oncology. In the preset data set, the classification error rate of the risk regression prediction was 30%, which is comparable to the apparent error rate (i.e. 33%) of the classification tree.

In a number of respects classification trees are decidedly different from traditional statistical methods for predicting class membership on a categorical dependent variable. They employ hierarchical modeling, with successive predictions being applied to particular cases, to sort the cases into homogeneous classes. Traditional methods use simultaneous techniques to make one and only one class membership prediction for each and every case. In other respects, such as having as its goal accurate foreboding, classification tree-based analysis is indistinguishable from traditional methods.

In epidemiology and in other prognostic research contexts in health fields involving complex data with many potential risk factors, the modern methods fall short of excelling

consistently those used commonly. It is yet premature to say if they have enough to commend themselves to become as accepted as the traditional methods. I conclude tentatively that tree-based analyses nevertheless have the potential power of providing complementary information and contributing to the interpretation of prognostication.

ACKNOWLEDGMENTS

I wish to thank Helena Isoniemi for kindly permitting me to use the kidney transplantation data collected at the Transplantation Unit of the Helsinki University Central Hospital, Helsinki, Finland, for illustration of prognostic modeling methodology. I am indebted to Terttu Kaustia for the English language revision.

CORRESPONDENCE TO

Dr Markku Nurminen, Department of Epidemiology and Biostatistics, Finnish Institute of Occupational Health, Topeliuksenkatu 41a A, FIN-00250 Helsinki, Finland. Phone: +358 9 4747 2408 Fax: +358 9 4747 2423 Email: markku.nurminen@ttl.fi

References

1. Miettinen OS. Evidence in medicine: invited commentary. Can Med Assoc J1998;158:215-21.
2. Isoniemi H, Nurminen M, Tikkanen M, Willebrand von, E, Krogerus L, Ahonen J, Eklund B, Höckerstedt K, Salmela K, Häyry P. Risk factors predicting chronic rejection of renal allografts. Transplantation 1994;106:68-72.
3. Valli H, Rosenberg PH, Kyttä J, Nurminen M. Arterial hypertension associated with the use of a tourniquet with either general or regional anaesthesia. Acta Anaesth Scand 1987;31:279-83.
4. Abu-Hanna A, Lucas PJF. Prognostic models in medicine [editorial]. Method Inform Med 2001;40:1-5.
5. Breiman LE, Davidoff F. Predicting clinical states in individual patients. Ann Intern Med1996;125:406-12.
6. Isoniemi H, Krogerus L, Willebrand von E, Taskinen E, Ahonen J, Häyry P. Histopathological findings in well-functioning, long-term renal allografts. Kidney Int 1992;41:155.
7. Cox DR. Regression models and life-tables [with discussion]. J R Stat Soc Ser B 1972;34:187-220.
8. Cox DR, Oakes D. Analysis of Survival Data. London: Chapman and Hal; 1984.
9. Nurminen M, Corvalán C, Leigh J, Baker G. Prediction of silicosis and lung cancer in the Australian labour force exposed to silica. Scand J Work Environ Health 1992;18:393-9.
10. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. New York (NY): Chapman and Hall; 1994.
11. Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
12. Lucas PJF and Abu-Hanna A. Prognostic methods in medicine. Artif Intell Med 1999;15:206-19.
13. Smith AFM. Discussion of "Fractional Bayes factors for model comparison" by A. O'Hagan. J R Stat Soc Ser B 1995;57:120-2.
14. Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-6.
15. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453-73.
16. Picard RR, Berk KN. Data splitting. Am Stat 1990;44:140-7.
17. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 1983;78:316-31.
18. Efron B, Tibshirani R. An Introduction to the Bootstrap. London: Chapman and Hall; 1993.
19. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values of confounding variables. Am J Epidemiol 1991;134: 895-907.
20. Greenland S, Finkle WD. A study of basic methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol1995;12:1255-64.
21. Venables WN, Ripley BD. Modern applied statistics with S-Plus, 3rd ed. New York (NY): Springer-Verlag; 1999.
22. McGullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman and Hall; 1989.
23. Marubini E, Valsecchi MG. Analysing Survival Data from Clinical Trials and Observational Studies. Chichester: John Wiley & Sons; 1995.
24. Breslow NE, Day NE. Statistical methods in cancer research. IARC scientific publications NO. 82, Lyon: International Agency for Research on Cancer; 1987.
25. S-PLUS 2000 Guide to Statistics. Seattle (WA): Data Analysis Products Division, MathSoft Inc.; 1999.
26. Clark LA, Pregibon D. Tree-based models. Chambers JM and Hastie TJ, editors. Statistical Models in S. New York, NY: Chapman and Hall; 1992. Ch. 9.
27. Aitchison TC, Sirel JM, Watt DC, MacKie RM [for the Scottish Melanoma Group]. Prognostic trees to aid prognosis in patients with cutaneous malignant melanoma. Br Med J1995;311:1536-9.
28. Ripley BD. Pattern recognition and neural networks. Cambidge: Cambridge University Press; 1996.
29. Greenland S. Summarization, smoothing, and inference in epidemiologic analysis. Scand J Soc Med1993;21:227-31.
30. Dannegger F. Tree stability diagnostics and some remedies for instability. Stat Med 2000;19:475-91.
31. Wyatt JC, Altman DG. Commentary: Prognostic models: Clinically useful or quickly forgotten? Br Med J 1995;311:1539-41.
32. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. Stat Med 1984;3:143-52.
33. Feinstein AR. Multivariable analysis: An antroduction. New Haven (CT): Yale University Press; 1996.
34. Copas JB. Regression, prediction and shrinkage [with discussion]. J R Stat Soc Ser B 1983;45:311-54.
35. van Houwelingen JC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med1995;14:1999-2008.
36. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan B, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 1997;9:107-38.
37. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent M-C. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. P Am Med Inform Assoc Sym 2000:156-60.

38. Lapuerta P, Azen PS, LaBree L. Use of neural networks in predicting the risk of coronary artery disease. Comput Biomed Res 1995;28:38-52.
39. Knuiman MW, Vu HT, Segal MR. An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. J Cardiovasc Risk 1997;4:127-34.
40. Duh MS, Walker AM, Ayanian JZ. Epidemiologic interpretation of artificial neural networks. Am J Epidemiol 1998;147:1112-22.
41. Duh MS, Walker AM, Pagano M, Kronlund K. Prediction and cross-validation of neural networks versus logistic regression: Using hepatic disorders as an example. Am J Epidemiol 1998;147:407-13.
42. Ioannidis JPA, McQueen PG, Goedert JJ, Kaslow RA. Use of neural networks to model complex immunogenetic associations of disease: Human leukocyte antigen impact on the progression of human immunodeficiency virus infection. Am J Epidemiol 1998;147:464-70.
43. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. Stat Med1998;17:2501-8.
44. Schwartzer G, Vach W. On the misuses of artificial neural networks for prognostic and diagnostic classification. Stat Med 2000;19:541-61.

Author Information

Markku Nurminen, Dr PH, Ph D
Finnish Institute of Occupational Health, University of Helsinki