STATISTICS IN MEDICINE
*Statist. Med.* **18**, 2529–2545 (1999)

# ASSESSMENT AND COMPARISON OF PROGNOSTIC CLASSIFICATION SCHEMES FOR SURVIVAL DATA

ERIKA GRAF*, CLAUDIA SCHMOOR, WILLI SAUERBREI AND MARTIN SCHUMACHER

*Institute of Medical Biometry and Medical Informatics, University of Freiburg, Stefan-Meier-Straße 26, D-79104 Freiburg, Germany*

## SUMMARY

Prognostic classification schemes have often been used in medical applications, but rarely subjected to a rigorous examination of their adequacy. For survival data, the statistical methodology to assess such schemes consists mainly of a range of *ad hoc* approaches, and there is an alarming lack of commonly accepted standards in this field. We review these methods and develop measures of inaccuracy which may be calculated in a validation study in order to assess the usefulness of estimated patient-specific survival probabilities associated with a prognostic classification scheme. These measures are meaningful even when the estimated probabilities are misspecified, and asymptotically they are not affected by random censorship. In addition, they can be used to derive $R^2$-type measures of explained residual variation. A breast cancer study will serve for illustration throughout the paper. Copyright © 1999 John Wiley & Sons, Ltd.

## 1. INTRODUCTION

The development of prognostic classification schemes to be used as clinical prediction rules is a point of major interest in many areas of clinical research such as cardiology, intensive care medicine or oncology. They are intended to provide a rational basis for making treatment or other clinical decisions, selecting homogeneous groups of patients for clinical trials, serving as stratification criteria in clinical trials and for conducting comparisons between institutions; their value and usefulness, however, is very much debated.[1] Problems in this area include the setting of adequate and accepted standards for the design, conduct and the statistical analysis of studies aimed to develop such prognostic classification schemes.[2,3] Besides this, there are additional intrinsic problems when a survival time or time to some event in the course of a disease is the endpoint of interest. Judging from the sparse empirical evidence, predictions of duration of survival tend to be rather inaccurate.[4–7] More precision may be achieved by using patient-specific survival probabilities as predictions to discriminate future survivors from failures.[8] Finally, censoring may cause technical problems in a survival context.[9]

In this paper, we review the statistical methods that are commonly used for the assessment and comparison of prognostic classification schemes. In survival analysis, they consist mostly of

* Correspondence to: Erika Graf, Institute of Medical Biometry and Medical Informatics, University of Freiburg, Stefan-Meier-Straße 26, D-79104 Freiburg, Germany. E-mail: egr@imbi.uni.freiburg.de

CCC 0277–6715/99/182529–17$17.50

Kaplan–Meier estimates of survival probabilities in groups of patients. Sometimes they are based on the likelihood function, on estimated regression coefficients of a regression model for survival data or on ROC methodology borrowed from the evaluation of diagnostic tests.[8–12] We will argue in favour of a measure of inaccuracy relating the estimated patient-specific survival probabilities to the observed outcome based on a suitable loss function.[7,9,13] The methods and associated problems are illustrated using data from a prospective, observational study in patients with node negative breast cancer[11] where disease-free survival is the endpoint of interest.

## 2. NOTATION AND DEFINITIONS

We assume that the prognosis of a particular group of patients can be described by the distribution of a positive random variable $T$ which represents the time from a well-defined starting point $t = 0$ to an event of interest occurring in the course of the disease. By $S(t) = P(T > t)$ we denote the marginal probability of being event-free up to time $t$. At time $t = 0$, which will usually be the time of diagnosis, start of treatment or time of randomization, we have information on a $p$-dimensional vector $X = (X^1, \ldots, X^p)$ of patient-specific covariates with values in some sample space $\mathcal{X}$. The covariates $X^1, \ldots, X^p$ are thought to have a potential influence on the event-free time $T$. Therefore they are often referred to as 'prognostic factors'. Given a vector of covariates $X = x$, observed at $t = 0$, the so-called 'patient-specific' probability of being event-free up to time $t$ is denoted by $S(t|X = x) = P(T > t|X = x)$. The aim of prognostic models in general can then be characterized as the attempt to estimate these patient-specific event-free probabilities as accurately as possible.

Often, these estimated probabilities are not given for every potential covariate combination $x \in \mathcal{X}$, but only for a finite number of groups of patients. We thus call a partition $\mathcal{X}_1, \ldots, \mathcal{X}_g$ of the sample space $\mathcal{X}$ a 'prognostic classification scheme' with 'risk strata' $\mathcal{X}_1, \ldots, \mathcal{X}_g$. The membership to a particular risk stratum can then be described by a one-dimensional covariate $\tilde{X}$ defined through $\tilde{X} = j$ if $X \in \mathcal{X}_j$, $j = 1, \ldots, g$. The corresponding estimated probabilities of being event-free up to time $t$ for patients in risk stratum $\mathcal{X}_j$ are denoted by $\hat{\pi}(t|\tilde{X} = j)$. The definition of risk strata and the estimated probabilities may have been derived from a careful statistical model building process; they can, however, also constitute a summary of subjective guesses from expert physicians or consensus meetings. The following questions arise immediately, and will be dealt with in this paper: How can one assess the accuracy of a given prognostic classification scheme? How can one compare different prognostic classification schemes with regard to their accuracy?

The terminology that may be found in the literature is extremely varied. For example reliability, precision, validity, discriminatory ability etc. have been used instead of accuracy. In the following, we will use the thoughtful terminology outlined in the monograph by Hand.[14]

## 3. EXAMPLE: PROGNOSTIC STUDY IN NODE NEGATIVE BREAST CANCER

For illustration, we will use data from a prospective study in node negative breast cancer conducted by the German Breast Cancer Study Group (GBSG).[11] From 1984 to 1989, a total of 662 patients were enrolled into the study, all of them having mastectomy and one cycle chemotherapy given perioperatively as standardized treatment. Age, menopausal status, tumour size, tumour grade, histologic tumour type, oestrogen receptor as well as progesteron receptor were recorded as prognostic factors. We restrict ourselves to 603 patients with complete data on the seven prognostic factors considered. The median follow-up of these is about five years; the

 

Table I. Prognostic study in node negative breast cancer: results from full Cox model and backward elimination. $\hat{\beta}$, estimated log-relative risk; RR, relative risk; $P$, $P$-value (full Cox model); $\hat{\beta}_{BE}$, estimated log-relative risk of selected factors after backward elimination

| Factor | | Number (per cent) | $\hat{\beta}$ | RR | $P$ | $\hat{\beta}_{BE}$ |
|---|---|---|---|---|---|---|
| Age | $\leqslant 40$ years | 61 (10) | 0 | 1 | 0·0004 | 0 |
| | $> 40$ years | 541 (90) | − 0·99 | 0·37 | | − 0·56 |
| Menopause | pre | 215 (36) | 0 | 1 | 0·0349 | — |
| | post | 388 (64) | 0·47 | 1·60 | | |
| Tumour size | $\leqslant 20$ mm | 281 (47) | 0 | 1 | 0·0002 | 0 |
| | $> 20$ mm | 322 (63) | 0·65 | 1·91 | | 0·59 |
| Oestrogen receptor | $< 20$ fmol | 270 (45) | 0 | 1 | 0·2102 | — |
| | $\geqslant 20$ fmol | 333 (55) | 0·27 | 1·31 | | |
| Progesterone receptor | $< 20$ fmol | 283 (47) | 0 | 1 | 0·5018 | — |
| | $\geqslant 20$ fmol | 320 (53) | − 0·14 | 0·87 | | |
| Tumour grade | I | 136 (23) | 0 | 1 | 0·0051 | 0 |
| | II | 325 (54) | 0·59 | 1·81 | | 0·59 |
| | III | 142 (23) | 0·93 | 2·55 | | 0·85 |
| Histological tumour type | solid | 300 (50) | 0 | 1 | 0·1221 | — |
| | ductal/lobular | 124 (20) | − 0·39 | 0·68 | | |
| | others | 179 (30) | − 0·31 | 0·73 | | |

endpoint of primary interest is disease-free survival which is defined as the time from treatment to the first of the following events: loco-regional recurrence; distant metastases; second cancer, and death. There have been 155 events observed so far; the Kaplan–Meier estimate of disease-free survival at five years is 0·733 (95 per cent confidence interval [0·694; 0·772]).

In a first step, the well-known Nottingham prognostic index (NPI)[15] is considered as a prognostic classification scheme according to the definition by Galea *et al.*[16] This index uses tumour size, lymph node stage and tumour grade as prognostic factors and is given by

$$\text{NPI} = (0{\cdot}2 \times \text{size}) + \text{lymph node stage} + \text{grade}.$$

Tumour size is given in centimetres, lymph node stage is 1 for node negative patients, 2 or 3 for node positive patients, and grade takes the values 1, 2 and 3. The risk strata are given by $\text{NPI} < 3{\cdot}4$, $3{\cdot}4 \leqslant \text{NPI} \leqslant 5{\cdot}4$ and $\text{NPI} > 5{\cdot}4$. Since the Nottingham prognostic index was originally developed for node negative and node positive patients, none of the patients in our study belongs to the 'high-risk' stratum $\text{NPI} > 5{\cdot}4$, so there are just two risk strata according to the NPI classification scheme in our prognostic study, $\mathscr{X}_1 : \text{NPI} < 3{\cdot}4$ ($n_1 = 266$) and $\mathscr{X}_2 : 3{\cdot}4 \leqslant \text{NPI} \leqslant 5{\cdot}4$ ($n_2 = 337$); the number of patients is given in brackets. In an attempt to develop a finer classification scheme for node negative patients, we fitted a Cox model[17] to our data, using all seven prognostic factors in a categorized form. The results given in Table I indicate that, in addition to tumour size and tumour grade, age and menopausal status exhibit a significant effect; patients younger than 40 years seem to have an increased risk of recurrence.

This Cox model can be seen as a prognostic classification scheme where the risk strata are defined through the entire set of combinations of categorized covariates. A simpler scheme may

 

be derived by categorizing the corresponding prognostic index (PI). PI is the sum of the covariate values of a particular patient, weighted by the corresponding estimated regression coefficients $\hat{\beta}$

$$\text{PI} = \hat{\beta}_1 X^1 + \cdots + \hat{\beta}_p X^p.$$

Categorizing PI, however, might still be too complicated for practical use, and some of the variables have only a small impact on PI. We therefore derived a simplified index from this model, using backward elimination at the 1 per cent level and rounded regression coefficients. The simplified index is defined as

$$\text{COX} = I(\text{age} \leqslant 40 \text{ years}) + I(\text{size} > 20 \text{ mm}) + \text{grade}$$

where $I(\cdot)$ denotes the indicator function being equal to 1 if '$\cdot$' holds true and 0 otherwise. The corresponding risk strata (with numbers of patients) are given by $\mathscr{X}_1$:COX $= 1, 2$ ($n_1 = 277$), $\mathscr{X}_2$: COX $= 3$ ($n_2 = 205$) and $\mathscr{X}_3$: COX $= 4, 5$ ($n_3 = 121$).

In addition, another prognostic classification scheme was derived from a classification and regression tree (CART) analysis. CART is a method designed to detect subgroups as homogeneous as possible within groups and as heterogeneous as possible between groups by recursively partitioning the covariate space.[18] Using CART in our data, we obtained four risk strata,[11] given as

| | | |
|---|---|---|
| $\mathscr{X}_1$: | grade 1 and age $\leqslant$ 60 years | ($n_1 = 78$) |
| $\mathscr{X}_2$: | size $\leqslant$ 20 mm and [(grade 2–3 and age $>$ 40 years) or (grade 1 and age $>$ 60 years)] | ($n_2 = 222$) |
| $\mathscr{X}_3$: | (age $\leqslant$ 40 years and grade 2–3) or (size $>$ 20 mm and age $>$ 60 years and grade 1) or (size $>$ 20 mm and grade 2–3 and oestrogen receptor $\leqslant$ 300 fmol) | ($n_3 = 284$) |
| $\mathscr{X}_4$: | size $>$ 20 mm and grade 2–3 and oestrogen receptor $>$ 300 fmol | ($n_4 = 19$). |

This leads to two relatively large 'medium' risk strata, a smaller 'low' risk stratum and a very small 'high' risk stratum in our study.

## 4. SOME *AD HOC* METHODS

In the context of survival analysis the method most commonly used for the assessment and comparison of prognostic classification schemes is the graphical presentation of Kaplan–Meier estimates[19] of the event-free survival probabilities in the corresponding risk strata. If a prognostic index derived from a Cox regression model is used, model-based estimated patient-specific event-free survival probabilities derived from that model may be presented instead.[20] Figure 1 shows the Kaplan–Meier estimates of disease-free survival in our study for the NPI (part (*a*)), for the simplified COX index (part (*c*)) and for the CART classification scheme (part (*d*)). For the full Cox model, estimated disease-free survival probabilities are shown for selected quantiles of the prognostic index (PI) in part (*b*) of Figure 1. Judging from visual inspection one gets the impression that the Cox model and the simplified Cox index yield a better separation than NPI and that there is additional improvement when applying the CART classification scheme. Obviously this is not surprising since NPI is the only classification scheme derived from external data. We will ignore this fact at this point and come back to it in the discussion.

The Kaplan–Meier estimates are often accompanied by *P*-values of the logrank test for homogeneity across risk strata.[21] Sometimes, a Cox model using dummy variates for the risk

 

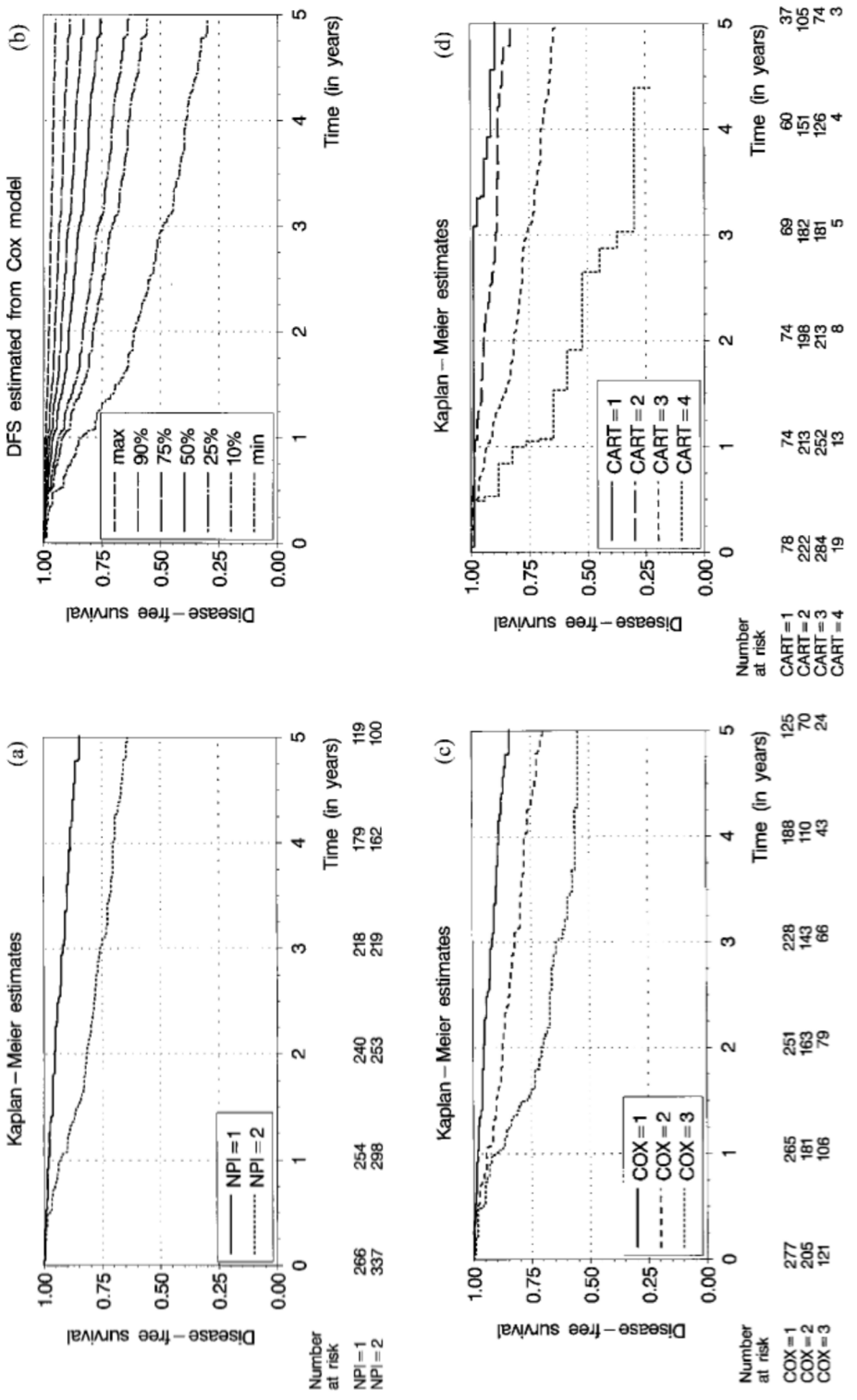


Figure 1. Prognostic study in node negative breast cancer: estimated probabilities of disease free survival. (*a*) Nottingham prognostic index (NPI), (*b*) full Cox model (PI), (*c*) COX index, (*d*) CART index

Table II. Prognostic study in node negative breast cancer: *ad hoc* measures

| | Pooled Kaplan–Meier | NPI | Full Cox model (PI) | COX index | CART index |
|---|---|---|---|---|---|
| *P*-value | — | 0·0001 | 0·0001 | 0·0001 | 0·0001 |
| −2 log-likelihood | 1856·1 | 1827·9 | 1807·1 | 1817·3 | 1802·9 |
| SEP | 1 | 1·559 | 1·624 | 1·550 | 1·710 |

strata is fitted and the log-likelihood and/or estimated relative risks of risk strata with respect to a reference are given. Recently, we have proposed a summary measure of separation[11] which is defined as

$$\mathrm{SEP} = \exp\left[ \sum_{j=1}^{g} \frac{n_j}{n} |\hat{\beta}_j| \right]$$

where $n_j$ denotes the number of patients in risk stratum $\mathscr{X}_j$ and $\hat{\beta}_j$ is the estimated log-hazard ratio or log-relative risk of patients in risk stratum $\mathscr{X}_j$ with respect to a baseline reference. In particular, we have used the baseline reference estimated in a Cox model where the dummy variates for risk strata were centred to have mean zero. SEP is the weighted geometric mean of 'absolute' relative risks between strata and baseline, 'absolute' meaning that 1/RR replaces RR for relative risks RR $< 1$. Often this model-based baseline reference turns out to be very similar to the estimated marginal distribution of $T$, that is, to the pooled Kaplan–Meier estimate $\hat{S}(t)$. Therefore SEP essentially compares risks within strata with the risk in the entire population. In fact the pooled Kaplan–Meier estimate has been used previously as baseline reference[11] although the model-based approach may be preferable for formal reasons.

Table II summarizes the results for the various *ad hoc* measures applied to the data of the prognostic study in node negative breast cancer. For all four prognostic classification schemes considered, the *P*-values of the logrank test are highly significant ($P \leqslant 0{\cdot}0001$). Thus, this measure does not prove to be particularly useful. There is some improvement in the log-likelihood for the NPI and some further improvement for the simplified Cox index and the full Cox model. The CART classification scheme shows the best result. A formal comparison, however, is hampered by the fact that the corresponding regression models are not nested. The summary measure SEP yields an average absolute risk with respect to baseline of about 1·56 for the NPI, of 1·55 for the simplified Cox index and of 1·62 for the full Cox model. With a value of 1·71 the CART classification scheme again shows the best performance.

## 5. MEASURES OF INACCURACY

A straightforward approach to the prediction of a time-to-event variable $T$ is to predict a time $\hat{T}(x)$ for a patient with $X = x$. For example, $\hat{T}(x)$ may be the median of an estimated survival curve $\hat{\pi}(t|\tilde{X} = j)$ for patients in risk stratum $\mathscr{X}_j$. $\hat{T}(x)$ might also be a subjective guess from an expert physician. For simplicity, we assume that the predicted time $\hat{T}(x)$ and the estimated event-free probabilities $\hat{\pi}(t|\tilde{X})$ in the risk strata $\mathscr{X}_1, \ldots, \mathscr{X}_g$ are given, fixed real-valued functions

 

of $\tilde{X}$, and that complete data $(T_i, X_i)$ are available for a 'test set' of $n$ patients $(i = 1, \ldots, n)$. Censoring is not considered so far. In this case, the mean square error of prediction

$$E_X[E[(T - \hat{T}(x))^2 | X = x]] = E[(T - \hat{T}(X))^2]$$

is an appropriate measure of inaccuracy which is readily estimated by

$$\frac{1}{n} \sum_{i=1}^{n} (T_i - \hat{T}(X_i))^2.$$

However it appears that point predictions of event-free times will almost inevitably give inaccurate and unsatisfactory results.[4-7] Thus, this approach will not be considered any further in this paper.

A second approach consists in predicting the survival or event status $Y = I(T > t^*)$ at a fixed time point $t^*$ by $\hat{Y}(x)$ for a patient with $X = x$. This could be done, for example, by setting $\hat{Y}(x) = 1$ if the estimated event-free probability $\hat{\pi}(t^* | \tilde{X} = j)$ exceeds a certain cut-off value $c$ for $x \in \mathscr{X}_j$, and 0 otherwise. In this case, the mean square error of prediction

$$E_X[E[(Y - \hat{Y}(x))^2 | X = x]] = E[Y \neq \hat{Y}(X)]$$

reduces to the expected misclassification or error rate. This quantity can be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq \hat{Y}(X_i)).$$

Other quantities like sensitivity and specificity or ROC curves can also be derived in a similar manner as in the context of diagnostic tests: the survival status $Y = I(T > t^*)$, 'survivor at $t^*$' or 'failure at $t^*$', corresponds to the disease status 'diseased' or 'not diseased'; the prediction $\hat{Y}(x)$, 'predicted survivor at $t^*$' or 'predicted failure at $t^*$' corresponds to the result of the diagnostic test 'classified as diseased' or 'classified as not diseased'. One problem with this approach is that at $t = 0$ when the prediction is made, the survival status at $t^*$ is not yet determined – it will evolve in the period from 0 to $t^*$ according to some stochastic mechanism. This is in contrast with the setting for diagnostic tests where the disease status of a patient is unknown but fixed at the time the diagnostic test is done. In the prognostic framework it is not adequate to judge patients as 'survivors at $t^*$' or 'failure at $t^*$', because at $t = 0$ the survival status at $t^*$ is not determined. In addition, the error rate has several disadvantages as a measure of inaccuracy,[5,14,23,24] one of them being that the information available is not fully exploited.

Thus, we consider a third approach which is based directly on the estimates of event-free probabilities $S(t^* | X = x)$ for patients with $X = x$. As outlined above, it is the aim of a prognostic classification scheme to provide estimated event-free probabilities $\hat{\pi}(t^* | \tilde{X} = j)$ for patients in risk stratum $\mathscr{X}_j$ $(j = 1, \ldots, g)$. These estimated probabilities may be used as predictions of the event status $Y = I(T > t^*)$. In order to determine the mean square error of prediction in this case, the observed survival or event status at $t^*$, $Y = I(T > t^*)$, has to be compared with the estimated probability $\hat{\pi}(t^* | \tilde{X})$, leading to

$$E_X[E[(Y - \hat{\pi}(t^* | \tilde{x}))^2 | X = x]] = E[(Y - \hat{\pi}(t^* | \tilde{X}))^2].$$

This quantity is known as quadratic score. Multiplied by a factor of two (omitted here for simplicity), it is equal to the expected Brier score, which was originally developed for judging the inaccuracy of probabilistic weather forecasts.[25-27] It should be noted that the misclassification

 

rate can be seen as some kind of coarsened version of the Brier score where $\hat{\pi}(t^*|\tilde{X})$ is replaced by $I(\hat{\pi}(t^*|\tilde{X}) > c)$. The expected Brier score may be interpreted as a mean square error of prediction when the estimated probabilities $\hat{\pi}(t^*|\tilde{X})$, which take values in the interval $[0, 1]$, are viewed formally as predictions of the event status at $t^*$, $I(T > t^*) \in \{0, 1\}$. The advantage of this approach is that estimated probabilities are used to predict the event status at $t^*$, which is a random variable at time $t = 0$. In the terminology of diagnostic tests this means that predictions are made in terms of 'predictive values' of a diagnostic test, that is, 'probabilities' of a positive or negative disease status, instead of classifying a patient as 'diseased' or 'not diseased'. It has been argued in the literature that even in the diagnostic setting the predictive value associated with the result of a diagnostic test is more relevant than the test result itself.[23,28,29] Hence to judge the quality of a diagnostic test, the Brier score, which measures average discrepancies between true disease status and estimated predictive values, may be preferred over the misclassification rate.[14,23,24,28,29]

We now return to the survival framework where the empirical version of the expected Brier score is readily defined as

$$\mathrm{BS}(t^*) = \frac{1}{n} \sum_{i=1}^{n} (I(T_i > t^*) - \hat{\pi}(t^*|\tilde{X}_i))^2.$$

In the extreme case where the estimated event-free probabilities are 0 or 1 for all patients – this corresponds to the assertion that the event-free status at $t^*$ can be predicted without error – $\mathrm{BS}(t^*)$ will be zero if $\hat{\pi}(t^*|\tilde{X}_i)$ coincides with the observed event status $I(T_i > t^*)$ for $i = 1, \ldots, n$; it will attain its maximum value of one only if the estimated event-free probabilities happen to be equal to $1 - I(T_i > t^*)$ for all patients. When the estimated event-free probabilities are taken to be constant, that is, equal for all patients say $\hat{\pi}(t^*|\tilde{X}_i) = \hat{\pi}(t^*)$, we have

$$\mathrm{BS}(t^*) = (\hat{\pi}(t^*) - \hat{S}(t^*))^2 + \hat{S}(t^*)(1 - \hat{S}(t^*))$$

where $\hat{S}(t^*)$ denotes the observed rate of event-free patients at $t^*$. In the absence of any knowledge about the disease under study, a trivial, constant prediction $\hat{\pi}(t^*) = 0{\cdot}5$ for all patients would be the most plausible approach. With a constant prediction of $\hat{\pi}(t^*) = 0{\cdot}5$, the empirical Brier score is equal to 0·25. The minimum value of $\hat{S}(t^*)(1 - \hat{S}(t^*))$ is attained for $\hat{\pi}(t^*) = \hat{S}(t^*)$.

There is a decomposition of the expected Brier score

$$E_X[E[(Y - \hat{\pi}(t^*|\tilde{x}))^2|X = x]]$$
$$= E_X[E[(\hat{\pi}(t^*|\tilde{x}) - S(t^*|x))^2|X = x]] + E_X[E[(Y - S(t^*|x))^2|X = x]]$$

that results in the conclusion that 'inaccuracy' can be split up into 'imprecision' and 'inseparability.[14] The second term is equal to $E_X[\mathrm{var}(Y|X = x)]$ and depends only on the event-free probabilities $S(t^*|X = x)$. Thus, it is equal for all possible prognostic classification schemes derived from the $p$-dimensional covariate vector $X$. The first term measures how well the estimated event-free probabilities $\hat{\pi}(t^*|\tilde{X})$ used as predictions correspond to the true event-free probabilities $S(t^*|X)$ and can therefore be seen as a measure of impression.

From this decomposition it can be seen that the expected Brier score takes its minimum value for $\hat{\pi}(t^*|\tilde{X}) = S(t^*|X)$, that is, when the true event-free probabilities are used as predictions. This property has been used as the definition of a so-called 'strictly proper scoring rule'.[14,23,27,29] The Brier score has a number of other interesting features. For example, there is also a useful

 

decomposition of its empirical version;[8,30] additional information can be gained from the distribution of individual contributions to the empirical Brier score.[8]

A more general definition of a measure of inaccuracy can be given by the introduction of a suitable loss function $L(T, \hat{\pi})$ leading to

$$\frac{1}{n}\sum_{i=1}^{n} L(T_i, \hat{\pi}(t^*|\tilde{X}_i))$$

as a general so-called 'scoring rule.[14,23,27] Clearly, $L(T, \hat{\pi}) = (I(T > t^*) - \hat{\pi}(t^*|\tilde{X}))^2$ gives the Brier score introduced above. Another obvious choice is to consider

$$L(T, \hat{\pi}) = -\{I(T > t^*)\log\hat{\pi}(t^*|\tilde{X}) + I(T \leqslant t^*)\log(1 - \hat{\pi}(t^*|\tilde{X}))\},$$

known as the 'logarithmic score', 'deviance' or 'negative log-likelihood'.[14,22,29,31] The empirical logarithmic score is given by

$$\mathrm{LS}(t^*) = -\frac{1}{n}\sum_{i=1}^{n}\{I(T_i > t^*)\log\hat{\pi}(t^*|\tilde{X}_i) + I(T_i \leqslant t^*)\log(1 - \hat{\pi}(t^*|\tilde{X}_i))\},$$

where we adopt conventions '$0 \times \log 0 = 0$' and '$1 \times \log 0 = -\infty$'. Hence $\mathrm{LS}(t^*)$ is equal to zero in the extreme situation where the estimated event-free probabilities $\hat{\pi}(t^*|\tilde{X}_i)$ are 0 or 1 for all patients and coincide with their observed event status, $I(T_i > t^*)$. It will attain infinity if the estimated event-free probability happens to be equal to $I(T_i \leqslant t^*)$ for at least one patient. Under the condition that the estimated event-free probabilities are constant and equal to $\hat{\pi}(t^*|\tilde{X}_i) = \hat{\pi}(t^*)$ for all patients, $\mathrm{LS}(t^*)$ attains its minimum value $-\{\hat{S}(t^*)\log\hat{S}(t^*) + (1 - \hat{S}(t^*))\log(1 - \hat{S}(t^*))\}$ for $\hat{\pi}(t^*) = \hat{S}(t^*)$. For the trivial prediction $\hat{\pi}(t^*) = 0{\cdot}5$ we get $\mathrm{LS}(t^*) = \log 2 = 0{\cdot}693$.

If we do not wish to restrict ourselves to one fixed time point $t^*$, we can consider the loss $L(T; \hat{\pi}(t|\tilde{X}))$ as a function of time $t$ and calculate the corresponding empirical loss for $0 \leqslant t \leqslant t^*$ to estimate the expected loss for all $t$. The loss can also be averaged over time, that is, for $t \in [0, t^*]$, by integrating it with respect to some weight function $W(t)$.[13] For example, with quadratic loss this approach yields an integrated version of the Brier score

$$\mathrm{IBS} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{t^*}(I(T_i > t) - \hat{\pi}(t|\tilde{X}_i))^2\,\mathrm{d}W(t) = \int_0^{t^*}\left\{\frac{1}{n}\sum_{i=1}^{n}(I(T_i > t) - \hat{\pi}(t|\tilde{X}_i))^2\right\}\mathrm{d}W(t).$$

Natural choices are $W(t) = t/t^*$ or $W(t) = (1 - \hat{S}(t))/(1 - \hat{S}(t^*))$ where $\hat{S}(t)$ denotes the estimated marginal survival function. For measures of inaccuracy based on other loss functions, an integrated version can be defined in a similar manner.

## 6. INCORPORATION OF CENSORING

Now we consider the situation of possibly censored time to event data. For each patient we observe $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leqslant C_i)$, where $T_i$ represents the time to the event of interest and $C_i$ the (hypothetical) time under observation ($i = 1, \ldots, n$). We assume that $T$ and the covariate $X$ are independent of $C$ – this is a common assumption of random censorship – and that $C$ is distributed according to $G(t) = P(C > t)$. For notational convenience we assume no ties between censored and uncensored observations.

 

For a fixed time point $t^*$, the contributions to the Brier score can be split up into three categories:

Category 1: $\tilde{T}_i \leqslant t^*$ and $\delta_i = 1$;
Category 2: $\tilde{T}_i > t^*$ ($\delta_i = 1$ or $\delta_i = 0$);
Category 3: $\tilde{T}_i \leqslant t^*$ and $\delta_i = 0$.

For the uncensored observations of category 1 the event occurred before $t^*$, and the event status at $t^*$ is equal to $I(T_i > t^*) = 0$; thus the contribution to the Brier score is $(0 - \hat{\pi}(t^*|\tilde{X}_i))^2$. In category 2 the observed event status at $t^*$ is equal to 1 since all of these patients are known to be event-free at $t^*$; the resulting contribution to the Brier score is $(1 - \hat{\pi}(t^*|\tilde{X}_i))^2$. For the censored observations of category 3 the censoring occurred before $t^*$ so that the event status at $t^*$ is unknown; thus their contribution to the Brier score cannot be calculated.

To compensate for the loss of information due to censoring, the individual contributions have to be reweighted in a similar way as in the calculation of the Kaplan–Meier estimator:[19] observations in category 1 get the weight $1/\hat{G}(\tilde{T}_i)$, those of category 2 get the weight $1/\hat{G}(t^*)$ and observations of category 3 get weight zero; here $\hat{G}(t)$ denotes the Kaplan–Meier estimate of the censoring distribution $G$, that is, the Kaplan–Meier estimate based on $(\tilde{T}_i, 1 - \delta_i)$, $i = 1, \ldots, n$. Divided by $n$, these weights sum up to 1, and the empirical Brier score under random censorship can be defined as

$$\mathrm{BS}^{\mathrm{c}}(t^*) = \frac{1}{n} \sum_{i=1}^{n} \{(0 - \hat{\pi}(t^*|\tilde{X}_i))^2 I(\tilde{T}_i \leqslant t^*, \delta_i = 1)\,(1/\hat{G}(\tilde{T}_i)) + (1 - \hat{\pi}(t^*|\tilde{X}_i))^2 I(\tilde{T}_i > t^*)\,(1/\hat{G}(t^*))\}.$$

Thus all observations including those censored before $t^*$ (category 3) contribute to the weights $1/(n\hat{G}(\tilde{T}_i))$ and $1/(n\hat{G}(t^*))$, but only those uncensored at $t^*$ (categories 1 and 2) contribute their estimated event-free probabilities $\hat{\pi}(t^*|\tilde{X}_i)$ to the calculation of $\mathrm{BS}^{\mathrm{c}}(t^*)$. If there is no censoring, $\mathrm{BS}^{\mathrm{c}}(t^*)$ reduces to $\mathrm{BS}(t^*)$, and, as $n$ increases, it converges to $E[(I(T > t^*) - \hat{\pi}(t^*|\tilde{X}))^2]$ for any censoring distribution $G$ provided $G(t^*) > 0$; for a detailed derivation and the handling of ties we refer to Graf.[32]

Again, $\mathrm{BS}^{\mathrm{c}}(t)$ can be calculated for all $t \leqslant t^*$, and an integrated version of the Brier score may be obtained by integrating $\mathrm{BS}^{\mathrm{c}}(t)$ with respect to a weight function $W(t)$. In this approach, censored observations will contribute their estimated event-free probabilities to the integrand up to the point $t$ where the censoring occurs. Censored observations can be treated in the same manner for loss functions other than the quadratic. The resulting measures of inaccuracy can be used to define measures of explained residual variation[13,33–38] for censored data in a straightforward and adequate way. If we use the Brier score at a fixed time point $t^*$, for example, we can calculate $\mathrm{BS}^{\mathrm{c}}(t^*)$ using $\hat{\pi}(t^*|\tilde{X}_i)$ as prediction of the event status for the patients in the various risk strata. In contrast to this, we can also calculate the empirical Brier score, say $\mathrm{BS}^{\mathrm{c}}_0(t^*)$, when $\hat{\pi}(t^*|\tilde{X}_i) = \hat{\pi}(t^*) = \hat{S}(t^*)$ is used as a prediction for all patients. Here $\hat{S}(t^*)$ is the Kaplan–Meier estimate at $t^*$ in the entire set of patients. The measure of explained residual variation is then defined as

$$R^2 = 1 - \mathrm{BS}^{\mathrm{c}}(t^*)/BS^{\mathrm{c}}_0(t^*).$$

Instead of the Brier score the integrated version of the Brier score or measures of inaccuracy based on other loss functions (for example, the deviance) can be used in the definition of $R^2$.

 

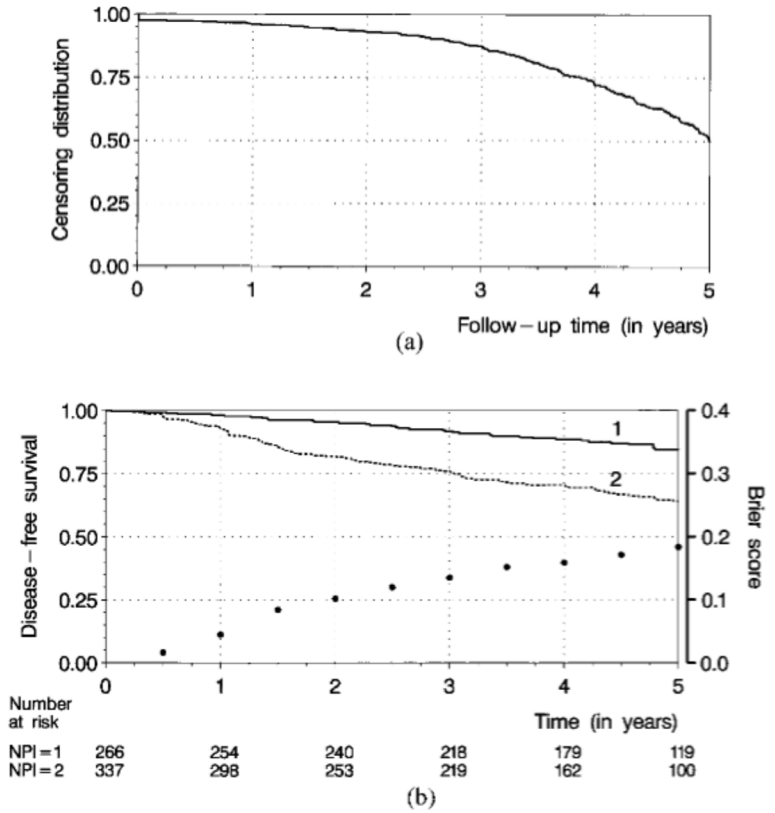


Figure 2. Prognostic study in node negative breast cancer: (a) Kaplan–Meier estimate $\hat{G}(t)$ of the censoring distribution $G(t) = P(C > t)$; (*b*) Kaplan–Meier estimate of disease-free survival for the Nottingham prognostic index (NPI), Brier score $BS^c(t)$ for the NPI and selected values of $t$

We calculated the Brier and the logarithmic score for the data of the prognostic study in node negative breast cancer. Figure 2(a) shows the estimated censoring distribution $\hat{G}(t)$ for these data. The Brier score for the NPI is displayed in Figure 2(b) as a function of time. For small values of $t^*$, say 0·5 or 1 year, the Brier score is relatively low. This means that predictions are very accurate for a short time period. For $t^* = 5$ the Brier score is equal to 0·184 which is not very much below 0·196, the value reached by the pooled Kaplan–Meier prediction $\hat{\pi}(t^*) = \hat{S}(t^*) = 0{\cdot}733$ for all patients; note that the Brier score is equal to 0·25 when the trivial prediction $\hat{\pi}(t^*) = 0{\cdot}5$ is made for all patients.

Table III summarizes the results of various measures of inaccuracy for the NPI, the full Cox model, the simplified COX index and the CART index. For all measures, the Cox model and the simplified COX index do better than the NPI, and some further improvement is achieved by the CART index. Relative to the prediction with the pooled Kaplan–Meier estimate for all patients, there is only a moderate gain of accuracy; the measure of explained residual variation just reaches 10·4 per cent for the best prognostic classification scheme.

 

Table III. Prognostic study in node negative breast cancer: measures of inaccuracy

| Measure of inaccuracy | Pooled Kaplan–Meier | NPI | Full Cox model (PI) | COX index | CART index |
|---|---|---|---|---|---|
| BS ($t^* = 5$) | 0·196 | 0·184 | 0·176 | 0·179 | 0·175 |
| IBS | 0·114 | 0·108 | 0·105 | 0·104 | 0·103 |
| LS ($t^* = 5$) | 0·580 | 0·549 | 0·532 | 0·538 | 0·529 |
| ILS | 0·372 | 0·351 | 0·343 | 0·341 | 0·336 |
| $R^2$ (BS($t^* = 5$)) | 0·0% | 6·1% | 10·3% | 8·7% | 10·4% |

## 7. DISCUSSION

The vast number of papers in the medical literature on prognosis and outcome prediction in various fields of clinical medicine makes it obvious that there is an urgent need for assessing the accuracy – or better the inaccuracy – of the numerous prognostic classification schemes proposed.[1] Although there are several *ad hoc* measures being frequently used there is no commonly accepted and sensible measure for such an assessment so far. Furthermore, if survival or a time-to-event variable has to be predicted there is some confusion about what quantities should or can be predicted and how censored observations should be dealt with.

Therefore, it is of central importance to recognize that the time-to-event itself cannot adequately be predicted.[4–7] The best one can do at $t = 0$ is to try to estimate the probability that the event of interest will not occur until $t^*$, given the available covariate information for a particular patient. Consequently, a measure of inaccuracy that is aimed to assess the value of a given prognostic classification scheme should compare the estimated event-free probabilities with the observed individual outcome. The *ad hoc* measures commonly used are only of limited value. In particular, ROC methodology borrowed from the evaluation of diagnostic tests often cannot adequately capture the features of a prognostic classification scheme. This applies also to the well-known *c*-index, an index of concordance[10,39,40] that enjoys some popularity.[41–46] The *c*-index has the interpretation of the area under the ROC curve for an artificially constructed 'diagnostic test' in the survival context, and hence the criticisms concerning the use of ROC curves[47–49] apply for these measures as well.

The properties of the Brier and the logarithmic score are well investigated and it is interesting to note that they have been also advocated for use in a diagnostic or classification setting.[14,23,29,50] In the field of prognosis, which also depends on the time horizon considered, they can be calculated and judged as functions of time. Summary measures of inaccuracy may be obtained by using loss functions integrated with respect to a suitable weight function.

A problem that has not been solved satisfactorily in the literature is the incorporation of censoring. We have presented a simple reweighting scheme which leads to quantities that do not depend on the censoring distribution asymptotically; the corresponding population values constitute meaningful quantities even when the prognostic classification scheme is grossly misspecified and whether or not the true distribution of the event-free time $T$ belongs to a particular class of models. In fact, their properties will hold even when predictions $\hat{\pi}$ are a continuous function of $X$. These advantages carry over to the resulting $R^2$-type measures where the problems of censoring and misspecification are still a matter of debate.[7,13,33–38]

 

A somewhat naive proposal that is sometimes used consists of simply omitting all censored cases from the data set.[51,52] In our breast cancer data, this leads to greater inaccuracy in terms of the Brier score as compared to those presented in Table III, thus we get a Brier score of 0·252 for the pooled Kaplan–Meier, 0·233 for the NPI, 0·227 for the COX index and 0·221 for the CART index. The corresponding $R^2$-values, however, are higher than those in Table III (0·076 for NPI, 0·102 for COX index, 0·123 for CART index), thus suggesting better values of explained residual variation than are actually achieved. This may be a feature of our specific data set. In the example we restricted ourselves to the period of up to 5 years, the median follow-up. At later times, the inaccuracy measures become greater in our data. This is because the Kaplan–Meier estimate approaches 0·5, a situation where accurate predictions are harder to make than initially, when almost everybody survives. The Brier score for the survival status at $t^* = 7$ years is even greater for the full Cox model than for the pooled Kaplan–Meier $\hat{S}$, yielding a negative explained residual variation measure. This may and indeed should happen for a model which is misspecified to such a degree that it yields predictions with a greater inaccuracy than the marginal distribution. It may well be that predictions from the full Cox model are too heavily model-based and that therefore they need not perform better than $\hat{S}$ for every single value of $t^*$. The other, simpler models have a lower Brier score than $\hat{S}$ at $t^* = 7$ years. Another reason may be that the percentage of censored observations increase in time, and this will surely affect the dispersion of the empirical Brier score. The question of how censoring in finite samples acts on the distribution of our measures of inaccuracy will be the subject of further research. Our current recommendation is to choose $t^*$ in a way that censoring is not too heavy (for example, the median follow-up time). We also prefer measures with integrated loss functions since they will reflect inaccuracy over an interval rather than just at one point in time. In addition, the corresponding empirical measures are likely to have lower dispersion, because censored observations contribute their estimated event-free probabilities $\hat{\pi}(t|\tilde{X})$ to the integrand until the censoring occurs.

A measure of separation like SEP may be used in situations where the proposed measures of inaccuracy cannot be calculated because a classification scheme is not supplemented by predictions of any kind (survival time, survival status or survival probability). Such a measure is constructed to assess by which amount survival within risk strata differs on average from survival in the entire population. The SEP measure presented here is based on the assumption of proportional hazards between risk strata. Within the proportional hazards model, the 'absolute' relative risk between two survival curves has the properties of a distance, on a multiplicative rather than an additive scale. Other distances between distributions not bound to a specific model might of course be considered to yield a broader class of separation measures.[14] For example, the average Kolmogorov–Smirnov statistic might be calculated for Kaplan–Meier estimates within risk strata compared to the pooled Kaplan–Meier curve. However, unless the proportional hazards assumption is grossly wrong, the interpretation as a relative risk is certainly an attractive feature of the SEP measure.

In this paper, several important topics have not been covered. The first is concerned with the overoptimism resulting when a measure of inaccuracy is calculated in the same data where the prognostic classification scheme is derived from. We have assumed that the estimated probabilities $\hat{\pi}(t|\tilde{X})$ of being event-free up to time $t$ have emerged from external sources. For our study in node negative breast cancer this is by no means true and only assumed for illustrative purposes. Actually, the full Cox model, the COX as well as the CART index have been derived from the same data set. Even for the Nottingham prognostic index that has been proposed in the literature[15,16] we estimated the event-free probabilities from our data set and used them as

 

predictions. To reduce the resulting overoptimism, cross-validation and resampling techniques may be employed in a similar way as for the estimation of error rates[53,54] or for the reduction of bias of effect estimates.[55] In a diagnostic setting, Linnet[29] has used a correction of the Brier score based on bootstrap resampling. For definitive conclusions, however, the determination of measures of inaccuracy in an independent test data set is absolutely necessary.[56]

The second topic not covered is the handling of missing covariate information. In our study in node negative breast cancer we have assumed complete information of the seven prognostic factors considered. This assumption has already reduced the data set from 662 to 603 patients. For future patients, however, one may wish to provide estimated event-free probabilities even if some factors have not been or could not be measured. To account for this, surrogate definitions of the prognostic classification scheme and the corresponding estimated event-free probabilities have to be derived for every missing pattern that can occur and/or is of particular interest. Obviously, these surrogate definitions and the occurrence of missing values will influence the measures of inaccuracy and have to be considered when assessing a prognostic classification scheme.

A third topic becomes relevant when a prognostic classification scheme should be applied to a new patient population where overall survival rates differ from those in the population for which the scheme was originally developed. In intensive care medicine, this calibration problem is usually called 'adjustment for the case mix'.[57] To deal with it, one would have to develop a procedure similar to the adjustment of the intercept term which is often done when a logistic egression model is to be used for prediction in a validation data set.[58] This is an issue for further research.

Of course, there are several other important topics that should be considered when assessing and comparing prognostic classification schemes, that have not been addressed here. We have, however, demonstrated that there are measures relating predicted event-free probabilities to observed individual outcome which constitute sensible tools for such an assessment. Moreover, these measures can account for censored observations is a simple and straightforward way. Finally, this paper may also contribute to the recent discussion on $R^2$-type measures of explained residual variation for censored survival or time-to-event data.[7,13,33–38]

ACKNOWLEDGEMENT

The authors would like to thank Robin Henderson from Lancaster University for stimulating discussions on prediction in the context of survival data.

## REFERENCES

1. Wyatt, J. C. and Altman, D. G. 'Commentary: prognostic models: clinically useful or quickly forgotten?', *British Medical Journal*, **311**, 1539–1541 (1995).
2. Simon, R. and Altman, D. G. 'Statistical aspects of prognostic factor studies in oncology', *British Journal of Cancer*, **69**, 979–985 (1994).
3. Laupacis, A., Sekar, N. and Stiell, I. G. 'Clinical prediction rules: a review and suggested modifications of methodological standards', *Journal of the American Medical Association*, **277**, 488–494 (1997).
4. Parkes, M. C. 'Accuracy of predictions of survival in later stages of cancer', *British Medical Journal*, **264,** 29–31 (1972).
5. Forster, L. A. and Lynn, J. 'Predicting life span for applicants to inpatient hospice', *Archives of Internal Medicine*, **148**, 2540–2543 (1988).

 

6. Maltoni, M., Pirovano, M., Scarpi, E., Marinari, M., Indelli, M., Arnoldi, E., Galluci, M., Frontini, L., Piva, L. and Amadori, D. 'Prediction of survival of patients terminally ill with cancer', *Cancer*, **75**, 2613–2622 (1995).
7. Henderson, R. and Jones, M. 'Prediction in survival analysis: model or medic?', *in* Jewell, N. P., Kimber, A. C., Ting Lee, M-L. and Withmore, G. A. (eds), *Lifetime Data: Models in Reliability and Survival Analysis*, Kluwer Academic publishers, Dordrecht, 1995.
8. Mackillop, W. J. and Quirt, C. F. 'Measuring the accuracy of prognostic judgements in oncology', *Journal of Clinical Epidemiology*, **50**, 21–29 (1997).
9. Henderson, R. 'Problems and prediction in survival-data analysis', *Statistics in Medicine*, **14**, 143–152 (1995).
10. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **15**, 361–387 (1996).
11. Sauerbrei, W., Hübner, K., Schmoor, C. and Schumacher, M. for the German Breast Cancer Study Group. 'Validation of existing and development of new prognostic classification schemes in node negative breast cancer', *Breast Cancer Research and Treatment*, **42,** 149–163 (1997). Correction, *Breast Cancer Research and Treatment*, **48,** 191–192 (1998).
12. van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **9**, 1303–1325 (1990).
13. Korn, E. L. and Simon, R. 'Measures of explained variation for survival data', *Statistics in Medicine*, **9**, 487–503 (1990).
14. Hand, D. J. *Construction and Assessment of Classification Rules*, Wiley, Chichester, 1997.
15. Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholoson, R. I. and Griffiths, K. 'A prognostic index in primary breast cancer', *British Journal of Cancer*, **45**, 361–366 (1982).
16. Galea, M. H., Blamey, R. W., Elston, C. E. and Ellis, I. O. 'The Nottingham Prognostic Index in primary breast cancer', *Breast Cancer Research and Treatment*, **22**, 207–219 (1992).
17. Cox, D. R. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
18. Breiman, L., Friedman, J. H., Ohlsen, R. A. and Stone, C. J. *Classification and Regression Trees*, Wadsworth, Belmont, 1984.
19. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
20. Andersen, P. K., Christensen, E., Fauerholdt, L. and Schlichting, P. 'Measuring prognosis using the proportional hazards regression model', *Scandinavian Journal of Statistics*, **10**, 49–52 (1983).
21. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S., Mantel, N., McPherson, K., Peto, J. and Smith, P. 'Design and analysis of clinical trials requiring prolonged observation of each patient. Part I', *British Journal of Cancer*, **34**, 585–612 (1976). Part II, *British Journal of Cancer*, **35**, 1–39 (1977).
22. Shapiro, A. R. 'The evaluation of clinical predictions', *New England Journal of Medicine*, **296**, 1509–1514 (1997).
23. Hilden, J., Habbema, J. D. F. and Bjerregard, B. 'The measurement of performance in probabilistic diagnosis III: methods based on continuous functions of the diagnostic probabilities', *Methods of Information in Medicine*, **17**, 238–246 (1978).
24. Habbema, J. D. F. and Hilden, J. 'The measurement of performance in probabilistic diagnosis IV: Utility considerations in therapeutics and prognostics', *Methods of Information in Medicine*, **20**, 80–96 (1981).
25. Brier, G. W. 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, **78**, 1–3 (1950).
26. Winkler, R. L. 'The quantification of judgement: some methodological suggestions', *Journal of the American Statistical Association*, **62**, 1105–1120 (1967).
27. Winkler, R. L. and Murphy, A. H. '"Good" probability assessors', *Journal of Applied Meteorology*, **7**, 751–758 (1968).
28. Spiegelhalter, D. J. 'Probabilistic prediction in patient management and clinical trials', *Statistics in Medicine*, **5**, 421–433 (1986).


 

29. Linnet, K. 'Assessing diagnostic tests by a strictly proper scoring rule', *Statistics in Medicine*, **8**, 609–618 (1989).
30. Yates, J. F. 'External correspondence: Decomposition of the mean probability score', *Organizational Behaviour and Human Performance*, **30**, 132–156 (1982).
31. Schmitz, P. I. M., Habbema, J. D. F. and Hermans, J. 'The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods', *Statistics in Medicine*, **2**, 199–205 (1983).
32. Graf, E. 'Explained variation measures for survival data', PhD thesis, University of Freiburg (in German), 1998.
33. Graf, E. 'Explained variation measures in survival analysis', *in* Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*, Vol. 2, Wiley, Chichester, 1998, pp. 1441–1443.
34. Graf, E. and Schumacher, M. 'An investigation on measures of explained variation in survival analysis', *Statistician*, **44**, 497–507 (1995).
35. Korn, E. J. and Simon, R. 'Explained residual variation, explained risk and goodness of fit', *American Statistician*, **45**, 201–206 (1991).
36. Schemper, M. 'The explained variation in proportional hazards regression', *Biometrika*, **77**, 216–218 (1990). Correction, **81**, 631 (1994).
37. Schemper, M. 'Further results on the explained variation in proportional hazards regression', *Biometrika*, **79**, 202–204 (1992).
38. Schemper, M. and Stare, J. 'Explained variation in survival analysis', *Statistics in Medicine*, **15**, 1999–2012 (1996).
39. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
40. Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. M., Waugh, R. A., Whalen, R. E. and Rosati, R. A. 'Predicting outcome in coronary disease: statistical models versus expert clinicians', *American Journal of Medicine*, **80**, 553–560 (1986).
41. Kong, D. F., Lee, K. L., Harrell, F. E., Boswick, J. M., Mark, D. B., Hlatky, M. A., Califf, R. M. and Pryor, D. B. 'Clinical experience and predicting survival in coronary disease', *Archives of Internal Medicine*, **149**, 1177–1181 (1989).
42. Hadorn, D. C., Draper, D., Rogers, W. H., Keeler, E. B. and Brook, R. H. 'Cross-validation performance of mortality prediction models', *Statistics in Medicine*, **11**, 475–489 (1992).
43. Marshall, G., Grover, F. L., Henderson, W. G. and Hammermeister, K. E. 'Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery', *Statistics in Medicine*, **13**, 1501–1511 (1994).
44. Lee, K. L., Woodlief, L. H., Topol, E. J., Weaver, D., Betriu, A., Col, J., Simoons, M., Aylward, P., Van de Werf, F. and Califf, R. M. 'Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: Results from an international trial of 41.021 patients', *Circulation*, **91**, 1659–1668 (1995).
45. Pelosio, P., Barbareschi, M., Bonoldi, E., Marchetti, A., Verderio, P., Caffo, O., Bevilacqua, P., Boracchi, P., Buttitta, F., Barbazza, R., Dalla Palma, P. and Gasparini, G. 'Clinical significance of cyclin D1 expression in patients with node-positive breast carcinoma treated with adjuvant therapy', *Annals of Oncology*, **7**, 695–703 (1996).
46. Farraggi, D., Simon, R., Yaskil, E. and Kramar, A. 'Bayesian neural network models for censored data', *Biometrical Journal*, **39**, 519–532 (1997).
47. McClish, D. K. and Powell, S. H. 'How well can physicians estimate mortality in a medical intensive care unit?', *Medical Decision Making*, **9**, 125–132 (1989).
48. Katz, D. and Foxman, B. 'How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability', *Epidemiology*, **4**, 319–326 (1993).
49. Detrano, R. 'Accuracy curves: An alternative graphical representation of probability data', *Journal of Clinical Epidemiology*, **42**, 983–986 (1989).
50. Hermans, J., Van Zomeren, B., Raatgever, W., Sterk, P. J. and Habbema, J. D. F. 'Use of posterior probabilities to evaluate methods of discriminate analysis', *Methods of Information in Medicine*, **20**, 207–212 (1981).
51. Burke, H. B. 'Artificial neural networks for cancer research: outcome prediction', *Seminars in Surgical Oncology*, **10**, 73–79 (1994).

 

52. Botacci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W. R., Macintyre, I. M. C., Duthie, G. S. and Monson, J. R. T. 'Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions', *Lancet*, **350**, 469–472 (1997).
53. Efron, B. 'Estimating the error rate of a prediction rule: improvement on cross-validation', *Journal of the American Statistical Association*, **78**, 316–330 (1983).
54. Efron, B. and Tibshirani, R. 'Improvement on cross-validation: the .632 + bootstrap method', *Journal of the American Statistical Association*, **92**, 548–560 (1997).
55. Schumacher, M., Holländer, N. and Sauerbrei, W. 'Resampling and cross-validation techniques: a tool to reduce bias caused by model building?', *Statistics in Medicine*, **16**, 2813–2827 (1997).
56. Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
57. Rowan, K. M., Kerr, J. H., McPherson, K., Short, A. and Vessey, M. P. 'Intensive Care Society's APACHE II Study in Britian and Ireland-II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method', *British Medical Journal*, **307**, 977–981 (1993).
58. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research*, *Vol. I*: *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon, 1980.

 