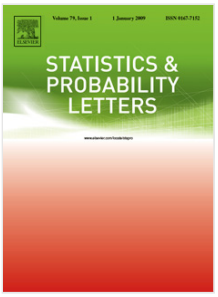
The use of ROC for defining the validity of the prognostic index in censored data

Petra Wolf, Georg Schmidt, Kurt Ulm



The use of ROC for defining the validity of the prognostic index in censored data

Petra Wolf, Georg Schmidt, Kurt Ulm

PII: S0167-7152(11)00067-8
DOI: 10.1016/j.spl.2011.02.021
Reference: STAPRO 5919

To appear in: *Statistics and Probability Letters*

Please cite this article as: Wolf, P., Schmidt, G., Ulm, K., The use of ROC for defining the validity of the prognostic index in censored data. *Statistics and Probability Letters* (2011), doi:10.1016/j.spl.2011.02.021

The Use of ROC for Defining the Validity of the Prognostic Index in Censored Data

Petra Wolf[*,1], Georg Schmidt[2], Kurt Ulm[1]

[1] Institut für Medizinische Statistik und Epidemiologie (IMSE) der Technischen Universität München, Ismaninger Straße 22, 81675 München, Germany
[2] I. Medizinische Klinik und Poliklinik der Technischen Universität München, Ismaninger Straße 22, 81675 München, Germany

Abstract

The validity of a diagnostic marker can be summarized using statistical measures either for the goodness of the fit like the deviance, measures of the explained variation like R^2 or the misclassification rate. Other intuitive measures are sensitivity and specificity in the case of binary response. In the absence of censored data the calculation of these measures is widely used. In the presence of censoring the estimation of time-dependent sensitivity and specificity is not well known. In this article we propose a new method of calculating ROC curves with censored data using the observed number of events and calculating the additional number of expected events for censored observations. The new method is illustrated with data for predicting mortality in patients surviving a myocardial infarction.

Key words: Discrimination, Nelson-Aalen estimator, Sensitivity, Specificity

1 Introduction

The identification of clinical markers to predict a defined event is an important topic in medicine. The event can be the diagnosis of a certain disease, the recurrences of a disease, the response to a certain therapy or mortality. The outcome of these markers may influence further diagnostic procedures or the selection of the appropriate therapy. Good examples can be found in oncology and cardiology, e.g. the PSA-testing for the diagnosis of prostate cancer (which is currently under debate) or the use of left-ventricular ejection fraction (LVEF) in predicting mortality after surviving an acute myocardial infarction.

The validity of these markers can be assessed with statistical measures either for the goodness of the fit, like the deviance, with measures of the explained variation like R^2 (Schemper and Henderson, 2000; O'Quigley and Xu, 2001) or the miss-classification rate. However the clinicians prefer measures like sensitivity and specificity, which can be summarized in ROC curves. If the defined event can be assessed within a short period of time, the use of ROC curves in order to present the sensitivity and specificity of a continuous marker is very popular. ROC curves can also be estimated using a combination of markers based on a statistical model. With these curves one can select a cut point in order to divide the marker or the prognostic index into two regions (e.g. positive and negative) based on the choice for sensitivity and specificity depending on the consequences. Therefore ROC curves provide an index of a test's ability to discriminate between alternative states of health.

[*] Corresponding author: e-mail: petra.wolf@tum.de, Phone: +49 89-41404348, Fax: +49 89-41404850

However in some studies there is a time lag between the measurement of the markers and the occurrence of the event. This is obvious when the event of interest is death. Some of the patients will die within the follow-up period, but there are always censored observations (patients still alive at the end of the observation period). In such a situation the ROC curves cannot be computed directly and it is difficult to demonstrate the validity of a prognostic marker. However there are some proposals in the literature how to calculate sensitivity and specificity in order to obtain ROC curves even in a study with failure time data (Heagerty et al., 2000; Heagerty and Zheng, 2005; Cai et al., 2006). Unfortunately these extensions are not well-known in the medical community, e.g. in the December issue of the New England Journal of Medicine in 2006 the following statement can be found: "Because standard methods do not exist for deriving ROC curves for time-to-event data ..." (Wang et al., 2006).

Besides the approaches by Heagerty et al. (2000), Heagerty and Zheng (2005) and Cai et al. (2006) calculating ROC curves, several related methods exist to evaluate the predictive accuracy: an often used and well-known measure is the concordance probability (c-index) (Harrell et al., 1996; Pencina and D'Agostino, 2004). Another possibility is the analytic calculation of the area under the ROC curve (AUC) (Chambless and Diao, 2006). If there is no censoring, c-index and AUC are identical and equate the Mann-Whitney statistics $P(X>Y)$. The Mann-Whitney statistic measures the probability that an observation drawn at random from population X (e.g. patients with disease) will exceed an observation drawn at random from population Y (e.g. patients without disease). However under random censorship there are quite differences between these measures: it is shown that Harrell's concordance index does not meet the Mann-Whitney statistic under censorship (Koziol and Jia, 2009). There are many extensions of Harrell's c-index, like methods using the inverse probability weighting (Liu and Jin, 2009; Uno et al., 2011) or with estimating the probability as a function of the Cox model (Gönen and Heller, 2005). These approaches got interesting characteristics and forms of application, but in order to use them properly one needs to know the objects they are measuring: the concordance index uses survival time as response whereas the AUC uses the binary endpoint event/no event at or up to a given time point. The Mann-Whitney parameter on the other hand compares two survival curves instead of comparing subjects with events and subjects without event.

Regarding the concordance probability also different definitions are possible: let T be the event time and Z a possible risk score. The conception in Harrell et al. (1996) and Pencina and D'Agostino (2004) is the evaluation of the probability $P(Z_1>Z_2|T_2>T_1)$. This definition can be linked to the AUC proposed by Heagerty and Zheng (2005) as a weighted area under the ROC curve. Gönen and Heller (2005) in contrast define a slightly different concordance probability: $P(T_2>T_1|Z_2\geq Z_1)$. With this definition it is possible to estimate the probability as a simple function of the Cox model. Another distinction must be made concerning the observation time: some measures quantify a probability at or until a given time, others are not interested in a particular time point and give an overall index for the whole course of time.

For clinical purpose three main aspects can be defined considering the prognostic ability of a marker: first of all the prognostic ability of a marker as addressed with the AUC or the c-index is of interest. The next step is the search for an adequate cut point to differentiate between subjects with high risk, who e.g. should get a treatment and patients with low risk for which a treatment is not essential. To enable a clinician to make such a decision, measures like sensitivity and specificity are useful tools. The information given by a c-index or an AUC alone is not sufficient. Third another useful information can be found in the time-dependence of the AUC. Is the prognostic ability the same over the complete observation time or is the marker only a good measure for some extent of time?

To answer these questions, we will present a new and simple method for calculating ROC curves for failure time data. Section 2 describes the estimation of time-dependent sensitivity and specificity in detail. In section 3 we compare the proposed method with the existing method of calculating time-dependent ROC curves by Heagerty et al. (2000). Section 4 demonstrates the method on data predicting mortality in patients surviving a myocardial infarction.

2 Methods

2.1 Notation

Assume a continuous marker Z has to be divided into low and high in order to use it for diagnostic or therapeutic purpose. If the disease status D is known for all subjects ($D = 0$ for no disease, $D = 1$ with disease) than sensitivity and specificity for a cut point z are defined as

$$\text{Se}(z) = \text{P}(Z > z \mid D = 1) \text{ and } \text{Sp}(z) = \text{P}(Z \leq z \mid D = 0).$$

For this definition we assume that the risk of the disease is positively correlated with the marker Z, otherwise the two regions $Z > z$ and $Z \leq z$ have to be exchanged.
In the situation of failure time data, the observation time t has to be taken into account. We use the following notation: T_i is the survival time for subject i and we are able to observe only the minimum of T_i and C_i where C_i is the independent censoring time. The follow-up time X_i is defined as the minimum of T_i and C_i ($X_i = \min(T_i, C_i)$) and δ_i is the censoring indicator; $\delta_i = 1$ represents an event at time X_i ($X_i = T_i \leq C_i$) whereas $\delta_i = 0$ indicates a censored observation ($X_i = C_i < T_i$).
To denote failure (disease) status of subject i at any time t, we use $D_i(t) = 1$ if $T_i \leq t$ and $D_i(t) = 0$ if $T_i > t$.

There are two proposals for the calculation of the sensitivity and specificity for time-to-event data (Heagerty and Zheng, 2005), the cumulative/dynamic and the incident/dynamic version:

a) *cumulative/dynamic (=C):*
The sensitivity for a given cut point z at time t is defined as:

$$\text{Se}^{\text{C}}(z,t) = \text{P}(Z > z \mid T \leq t)$$

The specificity is given by:

$$\text{Sp}^{\text{C}}(z,t) = \text{P}(Z \leq z \mid T > t)$$

With this definition all subjects will be used at any fixed time t. Subject i with survival time T_i will be counted as control up to time T_i and as a case afterwards.

b) *incident/dynamic (=I):*

$$\text{Se}^{\text{I}}(z,t) = \text{P}(Z > z \mid T = t)$$

$$\text{Sp}^{\text{I}}(z,t) = \text{P}(Z \leq z \mid T > t)$$

Using this definition, the number of subjects used will be decreasing with increasing observation time t. Subjects with a survival time T_i play the role of a control for time $t < T_i$ and then, for $T_i = t$ count as cases. But all subjects with $T_i < t$ do not contribute to the calculation of sensitivity/specificity at time t.
This definition is similar to the contribution of the observations available at time t to estimate certain parameters in the Cox-model using the partial likelihood method.
Both definitions have interesting properties. The cumulative/dynamic definition is useful if all the markers of interest are measured at baseline and one is interested in the prognostic properties of the markers up to time t.
The use of the incident/dynamic definition allows time-varying markers (Heagerty and Zheng, 2005).

In the following we are concentrating on the cumulative/dynamic definition. In the example considered the aim is to estimate sensitivity and specificity up to time t based on measurements obtained at baseline. Therefore we omit the superscript C for the cumulative/dynamic version of sensitivity and specificity.

2.2 Estimation

Using Bayes theorem we get

$$\mathrm{Se}(z,t) = \mathrm{P}(Z > z \mid T \le t) = \frac{\mathrm{P}(T \le t \mid Z > z) \cdot \mathrm{P}(Z > z)}{\mathrm{P}(T \le t)}$$

Consequently the specificity Sp(z, t) can be estimated as

$$\mathrm{Sp}(z,t) = \mathrm{P}(Z \le z \mid T > t) = \frac{\mathrm{P}(T > t \mid Z \le z) \cdot \mathrm{P}(Z \le z)}{\mathrm{P}(T > t)}$$

The problem in both formulas is related to the estimation of the probabilities $\mathrm{P}(T \le t \mid Z > z)$ and $\mathrm{P}(T > t \mid Z \le z)$ resp. $\mathrm{P}(T \le t)$ and $\mathrm{P}(T > t)$.
Heagerty et al. (2000) propose to use the Kaplan-Meier method to estimate theses probabilities. However the proportions of patients wit $Z > z$ are varying over time due to censorship and events. Therefore one can not use the result of the Kaplan-Meier curve for the whole sample. For censored observations it does not hold

$$P(T \le t) = P(T \le t \mid Z \le z) \cdot P(Z \le z) + P(T \le t \mid Z > z) \cdot P(Z > z)$$

with fixed proportions $\mathrm{P}(Z \le z)$ resp. $\mathrm{P}(Z > z)$, as should be valid under the law of total probability.

In order to obtain the ROC curve at time t one can estimate the expected number of events ($e_0(z,t)$ and $e_1(z,t)$) with $n_0(z)$ the number of observations wit $Z \le z$ and $n_1(z)$ those with $Z > z$ ($n_0(z) + n_1(z) = n$):

$Z \le z$: $e_0(z,t) = P(T \le t \mid Z \le z) \cdot n_0(z)$
$Z > z$: $e_1(z,t) = P(T \le t \mid Z > z) \cdot n_1(z)$

The sensitivity $\mathrm{Se}(z,t)$ can be estimated by $\frac{e_1(z,t)}{e_0(z,t) + e_1(z,t)}$ and the specificity accordingly.

The other option is to estimate the expected number of events at time t in using all observed events up to time t and calculate the additional number of events in using those who are censored before t.
For censored observations one has to estimate the probability for having an event between the censoring time c and time t ($t > c$) where the ROC curve should be calculated. This probability $\mathrm{P}(T \le t \mid T > c)$ can be estimated in using the hazard rates between c and t.
If the censoring occurs between t_{k-1} and t_k, the probability for having an event up to time t is given as

$$e = P(T \le t \mid T > c) = P(c < T \le t_k \mid T > c) + P(t_k < T \le t \mid T > c) = P(c < T \le t_k \mid T > c) + \frac{S(t_k) - S(t)}{S(c)}.$$

There are some possibilities to estimate the probability $P(c < T \le t_k \mid T > c)$: the exact calculation would be using the definition of a poisson process $P(c < T \le t_k \mid T > c) \approx (t_k - c)\lambda_{k-1}$. With this definition a person which is censored shortly after t_{k-1} has a higher probability of having an event than a person how is censored just before t_k. A slightly simplification is to assume that the event has occured in the middle between t_{k-1} and t_k, so that the probability $P(c < T \le t_k \mid T > c)$ can be estimated by $\frac{t_k - t_{k-1}}{2}\lambda_{k-1}$.
Under the assumption that the probability for a censored observation having an event is the same if it is censored at the beginning of the interval (t_{k-1}; t_k] or at the end of this interval or the interval is short, the probability for having an event reduces to

$$e = P(c < T \le t_k \mid T > c) + P(t_k < T \le t \mid T > c) = \frac{S(c) - S(t_k)}{S(c)} + \frac{S(t_k) - S(t)}{S(c)} = 1 - \frac{S(t)}{S(c)}$$
$$= 1 - \exp(-(\Lambda(t) - \Lambda(c))) \approx \Lambda(t) - \Lambda(c)$$

The approximation results as a first order Taylor approximation and is valid under the condition $(\Lambda(t) - \Lambda(c)) << 1$.

For this calculation, as an alternative to the Kalpan-Meier estimator, one can use the Nelson-Aalen estimator for the cumulative hazard rate

$$\Lambda(t) = \sum_{t_j \le t} \lambda(t_{(j)}) = \sum \lambda_j = \sum \frac{d_j}{n_j}$$

with d_j the number of events at time $t_{(j)}$ and n_j the risk set at $t_{(j)}$.

This calculation has to be done for all censored observations before t and has to be performed in both groups separately ($Z > z$ and $Z \le z$). With these probabilities together with the observed number of events one gets the expected number of events up to time t and one is able to construct the corresponding table to calculate sensitivity and specificity (Table 1).

Table 1. Data for calculating sensitivity and specificity at time t

time t	**with event**	**without event**	$\sum$
$\boldsymbol{Z > z}$	$E_1(t)$	$n_1 - E_1(t)$	$n_1(z)$
$\boldsymbol{Z \le z}$	$E_0(t)$	$n_o - E_0(t)$	$n_0(z)$
$\sum$	$E_1(t) + E_0(t)$	$n - (E_1(t) + E_0(t))$	n

with

$E_1(t) = \sum_{t_j \le t} d_{1j} + e_{1j}$ observed and expected number of events for $Z > z$ up to time t

$E_0(t) = \sum_{t_j \le t} d_{0j} + e_{0j}$ observed and expected number of events for $Z \le z$ up to time t

Based on this table one can easily calculate sensitivity Se(z, t) and specificity Sp(z, t) for all possible values of z and plot the ROC curve for selected time-points t:

$$\mathrm{Se}(z,t) = \frac{E_1(t)}{E_1(t) + E_0(t)}\ ;\ \mathrm{Sp}(z,t) = \frac{n_0 - E_0(t)}{n - (E_1(t) + E_0(t))}$$

One characteristic of a 'proper' ROC curve (Egan, 1975) is its monotonicity. The ROC curve is a monotone increasing function starting at (0,0) and ending at (1,1) (Pepe, 2003; Egan, 1975). Like with the proposed Kaplan-Meier estimator by Heagerty et al. (2000) there can be some situations where this assumption may be violated.

There is a proposal in the literature how to solve this problem in using the nearest neighbour estimator (NNE). This method was first described by Akritas (1994) and adopted for the given situation by Heagerty et al. (2000). All Kaplan-Meier curves near the cut point z are investigated. This approach guarantees monotonicity but it highly depends on the choice of the smoothing parameter, which can result in quite different estimates of sensitivity and specificity.

Another approach could be to calculate the ROC curves and apply afterwards a smoothing procedure to assure monotonicity.

One method available could be isotonic regression (Salanti and Ulm, 2005). If the monotonicity is failed than PAVA (**P**ooling **A**djacent **V**iolator **A**lgorithm) has to be applied. With this algorithm the two adjacent points with (Se(z, t), Sp(z, t)) and (Se(z+*1*, t), Sp(z+*1*, t)) are pooled together

$$\left(\hat{\mathrm{S}}\mathrm{e}(z,t),\hat{\mathrm{S}}\mathrm{p}(z,t)\right)=\left(\hat{\mathrm{S}}\mathrm{e}(z+1,t),\hat{\mathrm{S}}\mathrm{p}(z+1,t)\right)=\left(\frac{\mathrm{Se}(z,t)+\mathrm{Se}(z+1,t)}{2},\frac{\mathrm{Sp}(z,t)+\mathrm{S}p(z+1,t)}{2}\right)$$

in order to assure monotonicity.

A usual measure of the performance of the prognostic power of the marker Z is the AUC, the area under the ROC curve. In studies without censored data the AUC can be estimated in several ways, the Mann-Whitney-U statistic or the trapezoidal rule for integrating a function (Pepe, 2003). With censored data the trapezoidal rule seems to be appropriate.

With the AUC one is able to address two interesting questions.
a) Is there a difference between two markers (are two ROC curves different)?
b) Is there a change in the prognostic power of a marker over time?
For the comparison of two ROC curves their variances and covariances have to be calculated. Several methods are available for the comparison of two ROC curves if the event of interest is known (Pepe, 2003). In the case of censored data one can use bootstrap methods for calculating variances and confidence intervals. All other methods depend on the disease status which is not known for censored observations.

3 Comparison with existing methods

The existing method to calculate survival ROC curves via the Kaplan-Meier estimator (Heagerty et al., 2000) has one basic drawback. Despite the missing guarantee for monotonicity the major problem concerns the fact, that it doesn't satisfy the condition $0 \leq \mathrm{Se}(z,t), \mathrm{Sp}(z,t) \leq 1$. There can be situations where this condition is violated. Figure 1 presents data that illustrate this problem. These data describe the time to death of 863 kidney transplant patients with age as prognostic factor. A detailed description of the data can be found in Klein and Moeschberger (2003). In Figure 1 the ROC curve for age at time t=9 years is shown.

< insert figure 1 here >

For this data a cut point beneath an age of 25 leads to a sensitivity above 1. This peculiarity can be explained through the varying probabilities $P(Z > z)$ and $P(Z \leq z)$ as is shown in the Appendix.
Due to this inequality the simple approach with the Kaplan-Meier estimator together with the Bayes theorem doesn't provide a valid estimation of ROC curves in two respects. Besides the common problem with monotonicity particularly the problem with holding $0 \leq \mathrm{Se}(z,t), \mathrm{Sp}(z,t) \leq 1$ necessitates the complex calculation of ROC curves with an approach like the nearest neighbour estimation. A simple smoothing technique like the isotonic regression would fail in this setting because of the missing theoretical justification of this approach. This approach only works without censored observations.
To overcome this problem the nearest neighbour estimator was proposed. With this estimator monotonicity and the restriction $0 \leq \mathrm{Se}(z,t), \mathrm{Sp}(z,t) \leq 1$ is guaranteed. Though this causes another problem: how to choose the appropriate bandwidth for smoothing and what is the best bandwidth? There are no clear instructions in literature how to choose the bandwidth. In examining the course of the AUC over time there can be different best bandwidths at different time points. Simulation studies showed two different outcomes (results not shown): in some situations there is a huge difference in AUC for different bandwidth, but in other situations the resulting AUC for different bandwidths is not substantial different. But although the AUC was similar there were partly large differences in the values for sensitivity and specificity for a defined cut point z. A small bandwidth results in few steps in the ROC curve, which is

equivalent with few different values of sensitivity and specificity and a ragged curve with big steps. If one chooses a large bandwidth the curve is very smooth with many little steps and therefore many different values for sensitivity and specificity. Using the data of kidney transplant patients there results with a span for the nearest neighbor estimation between 0 and 0.2 with increments of 0.01 an AUC for predicting 9-years-survival between 0.663 and 0.695. An even more different result arises if one only uses the female: there results an AUC between 0.619 and 0.685 for predicting 9-years-survival.

The new proposed method for calculating ROC curves in the context of censored data overcomes these problems. Based on its definition the restriction $0 \leq Se(z,t), Sp(z,t) \leq 1$ is naturally achieved.
In this example the new ROC curve is similar to the one achieved with the Kaplan-Meier method but it can be seen in Figure 1 that there is a clear restriction to $0 \leq Se(z,t), Sp(z,t) \leq 1$.
A problem with monotonicity further persist but this can easy be solved with the introduced isotonic regression which is independent of any smoothing parameter and leads to an unique solution for sensitivity, specificity and the AUC.

4 Application

To demonstrate the new method, data from a study on patients who survived an acute myocardial infarction (MI) are analyzed.
This study has been used to derive a new prognostic marker (= DC) in order to predict mortality (Bauer et al., 2006). In this publication ROC curves were calculated to compare the prognostic impact of the new markers DC and AC with some of the established markers, e.g. LVEF and SDNN. In the publication the vital status at the end of an observation period of 2 years was used. There are only a limited number of patients with a follow-up less than two years. Therefore the effect of ignoring the censoring is fairly small.
In the meantime the follow-up period was extended. We want to demonstrate the prognostic impact of two markers LVEF and DC on the whole sample of n = 2343 patients.
The construction of the ROC curve will be first demonstrated using the established marker LVEF. The cut point used for defining the high risk population is $z = 30\%$. The patients with a LVEF of 30% or less are at increased risk. All analyses for this paper were done using `R 2.9.2` (R Foundation for Statistical Computing, Vienna, Austria). For the calculation of the expected number of events we used the exact formula via the poisson process as well as the simplification via $1-\exp(-(\Lambda(t)-\Lambda(c)))$. The AUC between these methods for LVEF and DC at 2 and 5 years only differed about 0.1%. This is due to the fact that in this study the event times are measured almost daily. Therefore the reported results correspond to the simpler formula which was more feasible to deal with for a bootstrap analysis.
Figure 2 shows the Kaplan-Meier curves for both subgroups (LVEF $\leq$ 30% resp. LVEF > 30%).

< insert figure 2 here >

Within five years 181 patients died (39 with LVEF $\leq$ 30%, 142 with LVEF > 30%) and 1022 observations were censored (40 with LVEF $\leq$ 30%, 982 with LVEF > 30%). The mortality rate at 5 years among patients with LVEF $\leq$ 30% was 37.91% and 7.64% among patients with LVEF > 30%.

The expected number of events at t=5 years, calculated as the sum of observed events and additional expected events for censored observations in both groups, results in 46.2 and 170.2 expected events compared to 39 and 142 observed events.
These values lead to the following estimates for the sensitivity and specificity of $z = 30\%$ at 5 years: the sensitivity is 21.3% and the specificity is 96.5%. In Figure 3 the ROC curve for LVEF at time point 5 years is shown.

< insert figure 3 here >

In order to investigate the change of the impact of LVEF over time the areas under the ROC curves are calculated and plotted against time t (Figure 4).

< insert figure 4 here >

It is easy to see that the impact of LVEF for predicting total mortality is very high within the first 6 months after the MI with a value of about 0.80. Afterwards the AUC drops down to values between 0.70 and 0.75.
The new marker DC is also a very powerful predictor. The ROC curve for DC at 5 years is shown in Figure 5. In order to compare both markers the area under each ROC curve is calculated and plotted over time. In Figure 6 the AUCs for both markers are shown. In this graph one can see the change of the impact of both markers over time. LVEF seems to be more powerful within the first 1.5 years. Afterwards there seems to be a slight advantage for DC. After 2.5 years both markers yield approximately the same AUC with values between 70% and 75%.

< insert figure 5 here >
< insert figure 6 here >

We used bootstrap methods to test the difference in AUC between the two markers at time points 2 and 5 years. The mean AUC difference over the 100 bootstrap samples was calculated, and the central 95% of the differences used as a confidence interval. The 95% confidence intervals for the differences were [-0.097; 0.0267] at 2 years and [-0.089; 0.0179] at 5 years. By the bootstrap confidence interval the AUCs from LVEF and DC at 2 and 5 years were not significantly different.

5 Discussion

With this article a new method of calculating sensitivity and specificity for censored data using the Nelson-Aalen estimator is introduced. When model development is intended to find important prognostic factors the key aspect is how to evaluate the predictive validity. A significant variable in a statistical model alone implies no reliable information about the additive prognostic value. An established discrimination measure is the area under the ROC curve. The AUC has a straightforward interpretation and in studies without censored data the ROC is a standard method. However, as mentioned in the introduction, there is no standard approach for deriving ROC curves for time-to-event data or more generally for the evaluation of the discriminating ability of prognostic markers. There a some proposals in literature, but no one has gained the standard approach.
The method proposed provides a simple straightforward calculation of sensitivity and specificity and the AUC if censoring occurs. For estimating sensitivity and specificity the unknown vital status for censored data at time t is substituted by the expected number of events at time t, using the observed events and calculating the additional expected events for censored observations. To achieve monotonicity for the ROC curve isotonic regression was used after calculating the raw sensitivities and specificities. This method to assure monotonicity displays an intuitive and easy way and is independent of the choice of any smoothing parameter.
Recently a new index was proposed for evaluating the added predictive ability of a new marker, the integrated discrimination index (IDI) (Pencina et al., 2008; Chambless et al., 2011). This index is based on integrated sensitivity and specificity and has a close relationship to the logistic regression R-square (Pepe et al., 2008). Like the AUC, it can be interpreted as a corrected average sensitivity. The idea for this approach is to overcome the drawback of the AUC in evaluating the added predictive ability of a new marker: to substantially improve the AUC a new marker must have a very large independent influence on the outcome. For this purpose the IDI is a more sensitive index. The increase in AUC is often small between a model with and without the new marker. In these cases the IDI can help to decide on the usefulness of a new marker.

Though an attractive approach, which was presented in this article, there are some more topics which warrant additional research. For our example we used bootstrap to obtain confidence intervals and to test the ability of discrimination of the two markers. But to make strong inference it would be desirable to have an analytic estimator for the variance of the ROC curve and the area under the curve. For evaluating long time periods the use of bootstrap is computationally intensive and it provides no rapid solution for practical application.

Acknowledgements
We would like to thank the Editor and a reviewer for helpful comments and suggestions that improved the content of the paper.

Conflict of Interests Statement
The authors declare no conflict of interest.

Appendix

The law of total probability states:
Let $B_1, B_2, \ldots, B_n$ be mutually exclusive events with $\sum P(B_i)=1$ for all $i=1,\ldots,n$. Then

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i) .$$

In the context of ROC curves with censored data the law of total probability can be formulated as:

$$P(T \le t) = P(T \le t \mid Z \le z) \cdot P(Z \le z) + P(T \le t \mid Z > z) \cdot P(Z > z) .$$

But with censored observations and fixed probabilities $P(Z \le z)$ and $P(Z > z)$ this equation does not hold, as can be shown with the following data of kidney transplant patients. As an example we use as cut point the median of age (43 years). An extract of the data is shown in the following table:

	total			age ≤ 43			age > 43		
time	n	events	S(t)	n	events	S(t)	n	events	S(t)
1	863	0	1	425	0	1	438	0	1
2	861	1	860/861	424	1	423/424	437	0	1
3	860	1	859/861	423	0	423/424	437	1	436/437
5	859	0	859/861	423	0	423/424	436	0	436/437
7	857	2	859/861· 855/857	422	1	423/424· 421/422	435	1	436/437· 434/435
…									

The first event occurs at t=2 but the first censoring is at t=1. It holds:

$$P(T \le 2) =$$

$$= \frac{860}{861} \ne \frac{423}{424} \cdot \frac{425}{863} + 1 \cdot \frac{438}{863} =$$

$$= P(T \le 2 \mid age \le 43) \cdot P(age \le 43) + P(T \le 2 \mid age > 43) \cdot P(T \le 2 \mid age > 43).$$

References

Akritas M. G. (1994). Nearest neighbour estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22**, 1299-1327.

Bauer A., Kantelhardt J. W., Barthel P., Schneider R., Mäkikallio T., Ulm K., Hnatkova K., Schömig A., Huikuri H., Bunde A., Malik M., Schmidt G. (2006). Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. *Lancet* **367**, 1674-81.

Cai T., Pepe M. S., Zheng Y., Lumley T., Jenny N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7**, 182–197.

Chambless L. E. and Diao G. (2006). Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in medicine* **25**, 3474–3486.

Chambless L. E., Cummiskey C. P., Cui G. (2011). Several methods to assess improvement in risk prediction models: Extension to survival analysis. *Statistics in medicine*, **30**, 22-38.

Egan J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, London.

Gönen M. and Heller G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika Trust* **92**, 965–970.

Heagerty P. J., Lumley T., Pepe M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Heagerty P. J. and Zheng Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Harrell F. E., Lee K. L., Mark D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15**, 361–387.

Klein J. P. and Moeschberger M. L. (2003). *Survival Analysis: techniques for censored and truncated data*. Springer, New York.

Koziol J. A., Jia Z. (2009). The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data. *Biometrical journal* **51**, 467–474.

Liu X. and Jin Z. (2009). A Non-Parametric Approach to Scale Reduction for Uni-Dimensional Screening Scales. *The International Journal of Biostatistics* **5**, Article 7.

Pencina M. J. and D'Agostino R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine* **23**, 2109–2123.

Pencina M. J., D'Agostino R. B., D'Agostino Jr. R. B., Vasan R. S. (2008). Evaluating the added predicitve ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157-172.

Pepe M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, New York.

Pepe M. S., Feng Z., Gu J. W. (2008). Comments on 'Evaluating the added predicitve ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). *Statistics in Medicine* **27**, 173-181.

Salanti G. and Ulm K. (2005). A non-parametric framework for estimating threshold limit values. *BMC Medical Research Methodoogyl*. **5:**36.

Schemper M. and Henderson R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56**, 249–255.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, n/a. doi: 10.1002/sim.4154.

O'Quigley J. and Xu R. (2001). *Explained variation in proportional hazards regression*. In: Handbook of Statistics in Clinical Oncology (ed J. Crowley), Marcel Dekker, New York, 397–409.

Wang T. J., Gona P., Larson M. G., Tofler G. H., Levy D., Newton-Cheh C., Jacques P. F., Rifai N., Selhub J., Robins S. J., Benjamin E. J., D'Agostino R. B., Vasan R. S. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *The New England Journal of Medicine* **355**, 2631–2639.

Figure Legends

Figure 1 ROC curves for 9-years-survival after kidney transplantation with age as prognostic marker. The solid line indicates the curve calculated with the Kaplan-Meier estimator and the dashed line indicates the new estimation with the Nelson-Aalen estimator.

Figure 2 Kaplan-Meier curve of mortality stratified by risk, according to left-ventricular ejection fraction (LVEF).

Figure 3 ROC curve for prediction of total mortality by LVEF at 5 years.

Figure 4 AUC(t) for prediction of total mortality based on LVEF with pointwise 95% confidence intervals.

Figure 5 ROC curve for prediction of total mortality by DC at 5 years.

Figure 6 AUC(t) for prediction of total mortality based on the two markers LVEF and DC.

Figure1

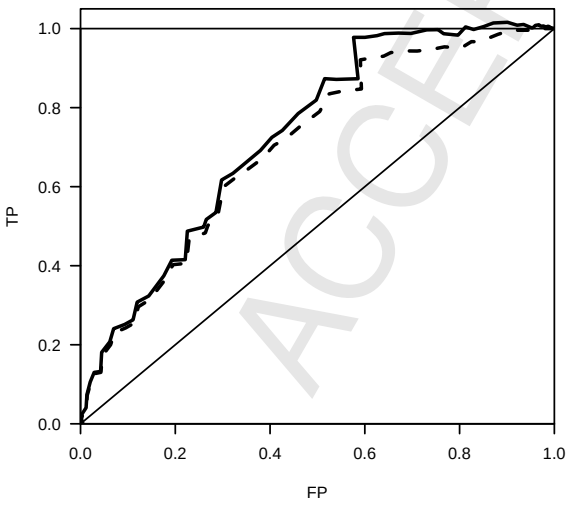
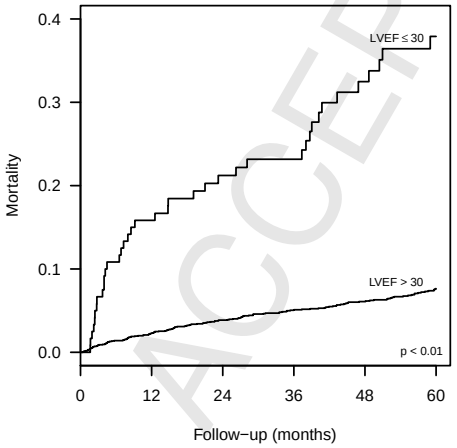


Figure2



LVEF ≤ 30	deaths	0	19	25	27	35	39
	n	120	100	82	70	52	41
LVEF > 30	deaths	0	50	84	106	122	142
	n	2223	2156	1912	1592	1339	1099

Figure3

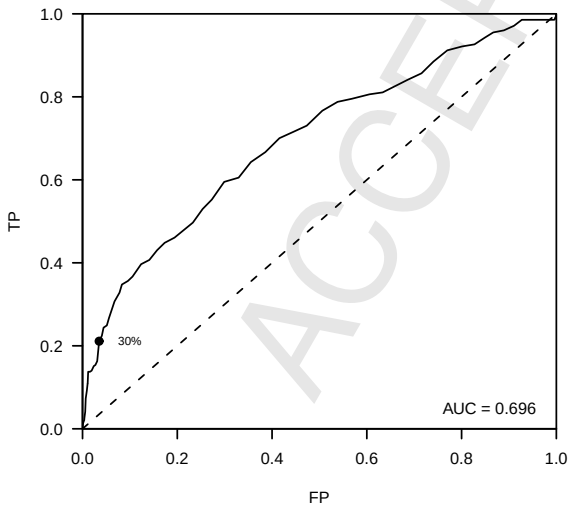


Figure4

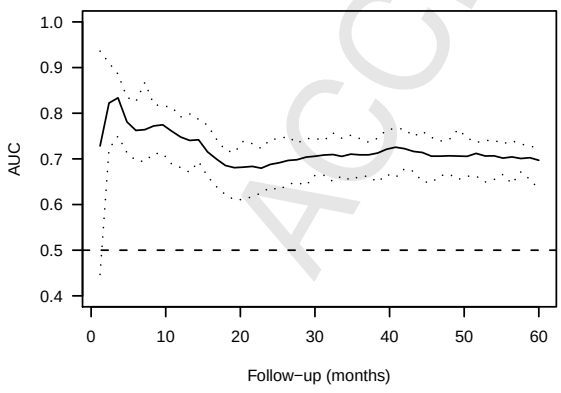


Figure5

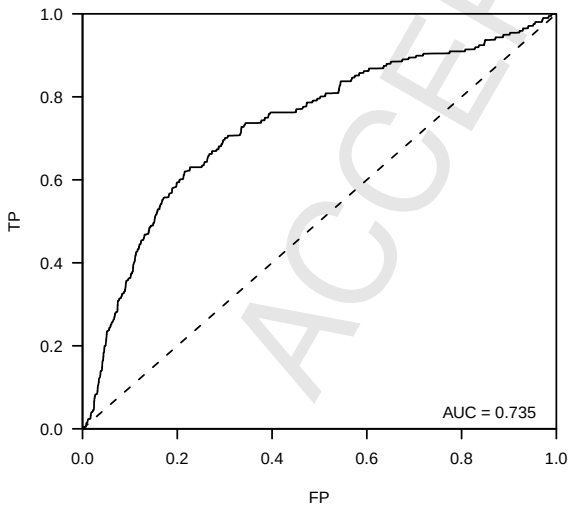


Figure6