An original approach to better evaluate the capacity of a prognostic marker using published survival curves

Etienne Dantan[1], Christophe Combescure[2], Marine Lorent[1], Joanna Ashton-Chess[3], Pascal Daguin[4], Jean-Marc Classe[5], Magali Giral[4,6] and Yohann Foucher[1]

[1] Department of Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences, EA 4275, Nantes University, Nantes, France

[2] CRC & Division of Clinical Epidemiology, Department of Health and Community Medicine, University of Geneva and University Hospitals of Geneva, Geneva, Switzerland

[3] TcLand Expression, Nantes, France

[4] Institute of Transplantation, Urology and Nephrology (ITUN), CHU Nantes, and INSERM U1064, Nantes, France

[5] Department of Surgical Oncology, Institut de Cancérologie de l'Ouest - Cancer Center René Gauducheau, Nantes, France

[6] CIC Biotherapy, CHU Nantes, France

Abbreviations:

ARE, acute rejection episode; HR, high risk; LR, low risk; MIG, monokine induced by interferon-γ; MR, medium risk; NHR, number of high risk individuals; NLR, number of low risk individuals; NMR, number of medium risk individuals; NPV, negative predictive value; PPV, positive predictive value

Corresponding author:

Etienne Dantan, Department of Biostatistics, Pharmacoepidemiology and Subjective Measures in Health Sciences, EA 4275, Nantes University, 1 rue Gaston Veil, 44035 Nantes, France (Etienne.Dantan@univ-nantes.fr).

Conflict of Interest/Financial Disclosure:

The authors have no financial disclosures of conflicts of interest to declare.

Abstract:

Objective: Predicting chronic disease evolution from prognostic marker is a key field of research in clinical epidemiology. However, the prognostic capacity of a marker is not systematically evaluated using the appropriate methodology. We proposed the use of simple equations to calculate time-dependent sensitivity and specificity based on published survival curves and other time-dependent indicators as predictive values, likelihood ratios and post-test probability ratios in order to re-appraise prognostic marker accuracy.

Study design and Setting: The methodology is illustrated by back-calculating time-dependent indicators from published papers presenting a marker as highly correlated with the time-to-event, concluding on the high prognostic capacity of the marker and presenting the Kaplan-Meier survival curves. The tools necessary to run these direct and simple computations are available online at http://www.divat.fr/en/online-calculators/evalbiom.

Results: Our examples illustrate that published conclusions about prognostic marker accuracy may be overoptimistic, thus giving potential for major mistakes in therapeutic decisions.

Conclusion: Our approach should help readers better evaluate clinical papers reporting on prognostic markers. Time-dependent sensitivity and specificity inform on the inherent prognostic capacity of a marker for a defined prognostic time. Time-dependent predictive values, likelihood ratios and post-test probability ratios may additionally contribute to interpreting the marker’s prognostic capacity.

Keywords:

Prognostic factor; Sensitivity; Specificity; Predictive values; Survival analysis, Likelihood ratios.

Running title:

Evaluation of marker prognostic capacity

Number of words in the abstract: 197

What is new?

- Time-dependent sensitivity and specificity as well as time-dependent predictive values can be find from published survival curves
- Time-dependent likelihood ratios and time-dependent post-test probability ratios are new indicators of prognostic marker accuracy
- The online application available at http://www.divat.fr/en/online-calculators/evalbiom allows back-calculating these time-dependent indicators
- The proposed time-dependent indicators should help readers better evaluate clinical papers reporting on prognostic markers.

In many therapeutic areas, predicting health events is a real challenge to improve the long-term medical management of patients affected by chronic disease. In characterizing a biological or pathological process, a surrogate marker may help to forecast a future event [1, 2]. A marker is often claimed as prognostic if it is significantly associated with the time-to-event distribution. This is valid at a population level, but this does not imply that the marker is a useful tool for individual decision-making. Rather, the clinical relevance of the marker should depend on its accuracy to predict a patient's evolution [3]. Therefore, clinical and biological prognostic markers, which can qualify a patient's likelihood of experiencing the event under consideration [3], are usually used to identify "at risk" patients who require more attentive follow-up or treatment adaptation. Understanding the term "prediction" as described in the literature can be baffling as this term can refer to the posterior estimation of a regression model, such as the survival probability 5 years post treatment, or to the prognosis of a future event up to a certain prognostic time, such as the death of a patient within the first 5 years post treatment. The present paper concerns this second definition. In the remainder of the document, the term "time-dependent" will be used for such prognosis up to a given prognostic time.

A widespread mistake when interpreting time-dependent data is confusing the notions of "prediction" and "correlation" [4]. The distance(s) between survival curves, the corresponding hazard ratio and the associated p-value are often presented as the most popular indicators of predictive capacity. However, p-values only demonstrate that the relationship is not a result of sample-to-sample fluctuation. Moreover, in a diagnostic context, Pepe et al. [5] demonstrated that the magnitude of an odds ratio has to be huge in order for it to inform on predictive capacity. Ware [6] pursued this line of investigation in a prognostic context and reported that the hazard ratio is not synonymous of prognostic capacity. According to Spruance et al. [7], the hazard ratio between patients at high and low risk of experiencing an event is also often wrongly interpreted since it only reflects the

magnitude of the changes in risk. Nevertheless, they demonstrated that the probability of experiencing the event sooner in the high risk group compared to the low risk group is equal to the hazard ratio divided by the hazard ratio plus one. Thus, the hazard ratio has to be massive to obtain a probability close to one, i.e. the reference value, qualifying the prognostic abilities of a marker on a time-scale change. It is quite common to find that markers, defined by authors as prognostic, are in fact only correlated with the survival outcome. Such common presentation of results can lead to overoptimistic and even erroneous conclusions and consequently misinterpretations concerning the potential utility of the marker [8].

In contrast with the diagnostic context, sensitivity and specificity are not often used in prognostic studies. The reason for this may be the difficulty in calculating these values when dealing with the time-to-event censoring process involved in long term studies, where the study patients do not all have the same follow-up time. Although Heagerty et al. [9] have described time-dependent estimators of sensitivity and specificity, these have not been widely adopted. Some authors estimate sensitivity and specificity based on the patients who have a follow-up at least equivalent to the prognostic time [10-12]. In this way, all patients who are censored before the prognostic time, i.e. who do not have this minimum follow-up, are excluded from the analysis. This results in considerable selection bias, leading to an over-representation of patients with failure or patients who reached this minimum follow-up period.

As highlighted by Riley et al. [13], the improving of all aspects of prognosis research is necessary to give better clinically relevant evidence for clinicians and health policy makers. Over the past few years, new statistical methods have been developed to evaluate the prognostic capacity of markers. However, these methods are not systematically used in medical publications to justify conclusions relating to the prognostic capacity of the reported markers. Back-calculation may be a solution to re-appraise the

conclusions reported in the published literature. For instance, Simel et al. [14] proposed an approach to calculate a posteriori sensitivity, specificity and likelihood ratios when the odds ratio and marginal numbers of a contingency table are the only information available in a published paper focusing on a diagnostic test. The aim of the present paper is to enable readers to reinterpret the prognostic capacity of markers based on survival curves already published in the literature. Our proposed approach would enable the reader to determine the marker's true prognostic capacity. We use simple equations to calculate time-dependent sensitivity and specificity from survival curves. Other time-dependent indicators, such as predictive values, likelihood ratios or post-test probability ratios are also described. We additionally provide simple illustrations in order to interpret the true meaning of results relating to the prediction of long-term outcome.

METHODS

Available information on published survival curves

The large majority of papers reporting on prognostic markers illustrate the results by plotting the non-adjusted Kaplan-Meier survival curves. Figure 1 illustrates the case of two survival curves for High Risk (HR) and Low Risk (LR) groups defined for a binary marker. Here NHR and NLR are respectively the numbers of individuals at baseline classified in the HR and LR groups. $SHR(t)$ and $SLR(t)$ are the corresponding survival probabilities at time *t* for these same groups. If more than two marker-based groups are studied, the binary assumption of the prognostic test no longer meets the definition of traditional sensitivity and specificity in which the outcome is binary, e.g. presence/absence of the disease. Nevertheless, in this context it is straightforward to merge groups, such as the HR and Medium Risk (MR) groups. If the $SMR(t)$ and NMR represent the survival probability at time *t* and the number of individuals at baseline in the MR group, respectively, this new HR group can be characterized by a survival that is equal to $(NHR * SHR(t) + NMR *$

$SMR(t))/((NHR + NMR))$. The MR group can also be merged with the LR group, which results in the calculation of two pairs of sensitivity and specificity values, one based on a strict and the other based on a lenient marker threshold.

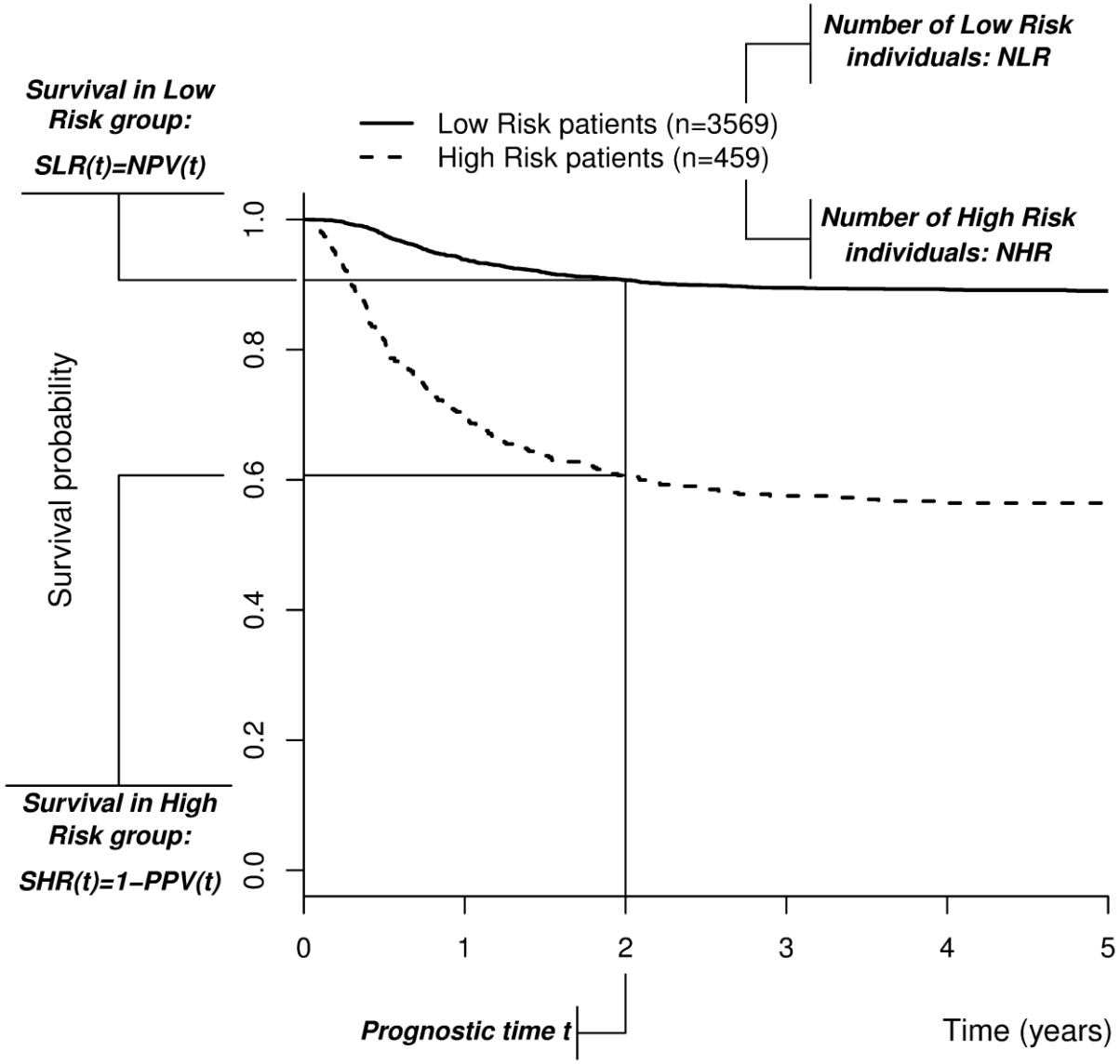


Figure 1. Available Information provided by most Kaplan-Meier Survival Curves. The Four Key Parameters to be extracted are: the Baseline Number of HR Patients (NHR), the Baseline Number of LR Patients (NLR), the Survival at Time t in the HR Group (SHR), the Survival at Time t in the LR Group (SLR).

Time-dependent sensitivity and specificity

The prognosis is made up to the prognostic time *t*. Then, $D(t)$ is the time-dependent indicator of the event with $D(t) = 1$ if the event occurs before *t* and $D(t) = 0$ otherwise. Based on the definitions of Heagerty et al. [9], the sensitivity at time *t* represents the proportion of patients who are correctly classified as HR among all patients who experience the event before time *t*, i.e. $Se(t) = P(HR|D(t) = 1)$. The specificity at time *t* represents the proportion of patients who are correctly classified as LR among all patients who do not experience the event before *t*, i.e. $Sp(t) = P(LR|D(t) = 0)$. Using the Bayes theorem (demonstration provided in the appendix), time-dependent sensitivity and specificity for a prognosis up to time *t* can be easily calculated from survival at time *t* and the baseline numbers of individuals in the HR and LR groups:

$$Se(t) = \frac{\big(1 - SHR(t)\big) \times NHR}{\big(1 - SHR(t)\big) \times NHR + \big(1 - SLR(t)\big) \times NLR} \quad (1)$$

$$Sp(t) = \frac{SLR(t) \times NLR}{SHR(t) \times NHR + SLR(t) \times NLR} \quad (2)$$

Time-dependent predictive value

While sensitivity and specificity indicate the intrinsic qualities of the marker by indicating risk group probabilities based on the future event status, predictive values are useful indicators for practical clinical interpretation and decision making. In fact, the time-dependent positive and negative predictive values, denoted $PPV(t)$ and $NPV(t)$, are directly obtained from the survival at time *t*. The positive predictive value is the probability that HR patients will experience the event before time *t*, i.e. $PPV(t) = P(D(t) = 1|HR) = 1 - SHR(t)$. The negative predictive value is the probability that LR patients will not

experience the event before time t, i.e. $NPV(t) = P(D(t) = 0|LR) = SLR(t)$. The confidence intervals of each survival probability may sometimes be presented in the survival graphs or specified in the text for different time-points. Therefore, the corresponding confidence interval for both predictive values can be obtained directly.

Time-dependent predictive values can also be defined as functions of time-dependent sensitivity and specificity (demonstration provided in the appendix):

$$PPV(t) = \frac{Se(t) \times P(D(t)=1)}{Se(t) \times P(D(t)=1) + (1 - Sp(t)) \times (1 - P(D(t)=1))} \quad (3)$$

$$NPV(t) = \frac{Sp(t) \times (1 - P(D(t)=1))}{Sp(t) \times (1 - P(D(t)=1)) + (1 - Se(t)) \times P(D(t)=1)} \quad (4)$$

Regarding the previous equations, the main limitation of predictive values is their dependence on the population frailty, i.e. the event probability. Therefore, without the calculation of previous time-dependent sensitivity and specificity values, the conclusions about prognostic capacity from a specific study cannot be directly generalized to other populations where the event probability is different. This significant limitation of predictive values is well accepted in traditional diagnostic medicine but is always ignored in prognostic analyses.

Time-dependent likelihood ratios

In the diagnostic context, positive and negative likelihood ratios are well defined [15, 16] and often used in practice. To our knowledge, their definitions have never been adapted to a prognostic context. Positive and negative likelihood ratios for a prognosis up to time *t*, denoted $LikR^+(t)$ and $LikR^-(t)$, are respectively:

$$LikR^+(t) = \frac{P(HR|D(t)=1)}{P(HR|D(t)=0)} = \frac{Se(t)}{1-Sp(t)} \quad (5)$$

$$LikR^{-}(t) = \frac{P(LR|D(t)=1)}{P\left(LR\middle|D(t)=0\right)} = \frac{1-Se(t)}{Sp(t)} \qquad (6)$$

If time-dependent likelihood ratios are close to 1, then the classification rule tends not to be informative of the future event. The higher the positive likelihood ratio, the more the HR group probability is associated with the occurrence of the event before time *t*. Conversely, the lower the negative likelihood ratio, the more the LR group probability is associated with the absence of the event before time *t*. Given the thresholds used in diagnostic evaluations, positive and negative likelihood ratios close to 10 and 0.1, respectively, indicate a useful marker for a prognosis up to the time *t* [15]. Since time-dependent likelihood ratios are derived from sensitivity and specificity at time *t*, they are independent of the event probability and thus represent intrinsic characteristics of the marker's prognostic capacity.

Time-dependent post-test probability ratios

Let $PT^{+}(t)$ and $PT^{-}(t)$ be the positive and negative post-test probability ratios, respectively. Both quantities can be expressed according to the time-dependent likelihood ratios (demonstration provided in the appendix):

$$PT^{+}(t) = \frac{P(D(t)=1|HR)}{P(D(t)=0|HR)} = \frac{P(D(t)=1)}{P(D(t)=0)} LikR^{+}(t) \qquad (7)$$

$$PT^{-}(t) = \frac{P(D(t)=1|LR)}{P(D(t)=0|LR)} = \frac{P(D(t)=1)}{P(D(t)=0)} LikR^{-}(t) \qquad (8)$$

Similarly to diagnostic context, the time-dependent positive and negative likelihood ratios appear as a multiplicative coefficient between pre-test probability ratio and post-test probability ratio. The corresponding interpretations are straightforward. A patient classified in the HR group has a $PT^{+}(t)$ times greater risk of presenting the event before time *t* than after *t*. In contrast, a patient classified as LR has a $1/PT^{-}(t)$ times greater risk of

presenting the event after time *t* than before *t*. Unlike the likelihood ratio, these values must be interpreted with caution given their dependence on the population frailty.

Software

All of the methods described above can be performed using the online tools available at http://www.divat.fr/en/online-calculators/evalbiom.

APPLICATIONS

To illustrate the practical utility of the proposed methods, we have selected two papers: one in kidney transplantation and one in breast cancer. Nevertheless, to confirm our approach, we have also applied it to several other papers in various fields of medicine (results not shown). For this purpose we only chose papers in which *i)* the markers were highly correlated with the time-to-event with small p-values, *ii)* the authors concluded on the high prognostic capacity of the marker and *iii)* the NHR, NLR , $SHR(t)$ and $SLR(t)$ were available. This selection of papers does not achieve the standards of a meta-analysis, the purpose of this selection is simply to highlight the usefulness of the proposed approach.

Prognosis of an acute rejection episode after kidney transplantation

In the study by Hauser et al. [17], the outcome was the time between the kidney transplantation and the first Acute Rejection Episode (ARE). The HR group was made up of 20 patients presenting an increase in urinary monokine induced by IFN-γ (MIG) above 436 pg/ml (positive MIG). The LR group comprised 49 patients presenting no increase in urinary MIG (negative MIG). The survival curves differed significantly between the two groups ($p<0.0001$). More precisely, ARE survival probability at 40 days post transplantation was 0.29 in the HR group and 0.99 in the LR group. The overall cumulative probability of

ARE beyond 40 days post-transplantation was 0.21. Applying equations 1 and 2, the sensitivity of a prognosis up to 40 days was 0.97 and the specificity was 0.89 (Table 1). For a shorter prognosis up to 10 days, the sensitivity appeared to be perfect but the specificity decreased to 0.76. Urinary MIG thus seemed to be a very sensitive and specific predictor of ARE after kidney transplantation. More precisely, a decision based on an increase in urinary MIG limited the detection of false negatives. At 40 days, less than 3% of patients were incorrectly classified as LR. Nevertheless, this was at the cost of a slightly lower specificity, i.e. around 11% of HR patients were incorrectly classified for a prognosis at 40 days.

Table 1. Description of the Results obtained from 2 Different Papers based on the Data Available in the Corresponding Papers and for the Purpose of this Report, Estimations Calculated using the Proposed Equations.

Data extracted from the paper						The estimations of the novel indicators							
Prognostic time t	NHR	SHR	NLR	SLR	p-value	$Se(t)$	$Sp(t)$	$PPV(t)$	$NPV(t)$	$LikR^{+}(t)$	$LikR^{-}(t)$	$PT^{+}(t)$	$PT^{-}(t)$
*Hauser et al. (2005)**													
10 days	20	0.76	49	1.00	<0.01	1.00	0.76	0.24	1.00	4.22	0.00	0.32	0.00
40 days		0.29		0.99		0.97	0.89	0.71	0.99	9.05	0.04	2.45	0.01
*Mook et al. (2010)***													
5 years	439	0.80	525	0.95	<0.01	0.77	0.59	0.20	0.95	1.86	0.39	0.25	0.05
10 years		0.72		0.87		0.64	0.59	0.28	0.87	1.57	0.60	0.39	0.15

* *Endpoint: Distant metastasis after surgery, Biomarker: Urinary monokine (MIG), HR group: MIG elevation above 436 pg/ml*

** *Endpoint: Acute rejection after renal transplantation, Biomarker: 70-gene MammaPrint signature, HR group: Poor prognosis group using the 70-gene MammaPrint signature*

For a prognosis up to 40 days, the PPV and NPV were estimated at 0.71 and 0.99, respectively. This illustrates the high accuracy of a negative test result in practice: patients without an increase in MIG had a less than 1% risk of acute rejection within the first 40 days post transplantation. The prognostic accuracy of a positive test result was also reasonable: patients with an increase in MIG had a 71% risk of acute rejection during the first 40 days post transplantation. As previously demonstrated, the use of urinary MIG as a biomarker in other populations, i.e. with different cumulative event probabilities, may lead to an unexpected PPV and NPV. Based on equations 3 and 4, Figure 2 illustrates this variation. In France for example, the cumulative probability of an ARE at 40 days post renal transplantation is around 9% (data from the DIVAT network, www.divat.fr). The PPV and NPV would be 0.47 and 0.99, respectively. Thus even though the NPV is always expected to be almost perfect, the PPV decreases. As a result, 53% of patients who tested positive for MIG would not have an ARE.

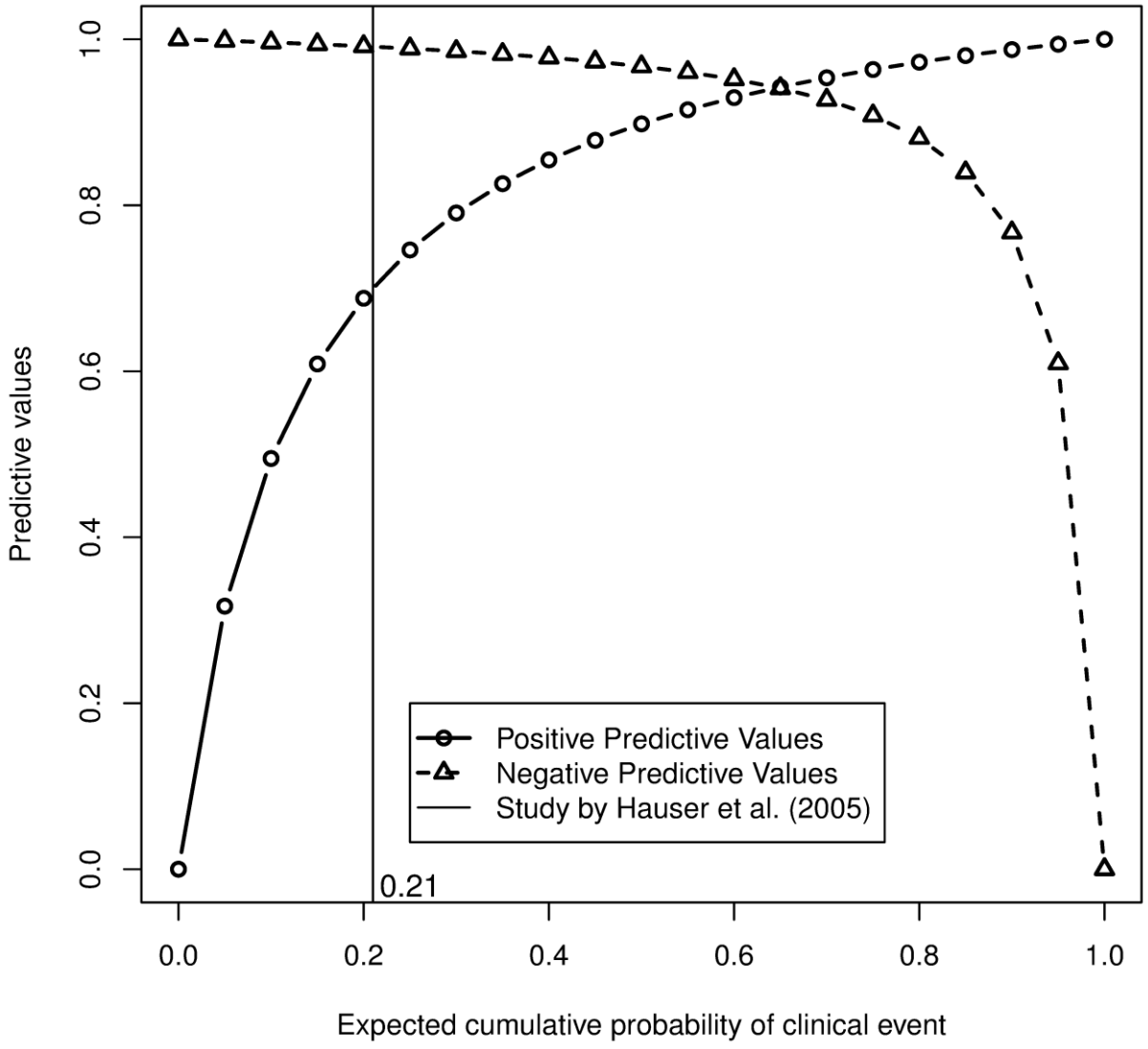


Figure 2. Positive and Negative Predictive Values regarding the Expected Cumulative Probability of Acute Rejection up to 40 days post transplantation in Kidney transplant Recipients. In the study by Hauser et al. (2005), the Probability of Acute Rejection before 40 days was 21%; the Positive and Negative Predictive Values were 71% and 99%, respectively.

At 40 days, the time-dependent positive and negative likelihood ratios were 9.05 and 0.04. Time-dependent likelihood ratios were thus not too different from the expected values of 10 and 0.1, respectively. The urinary MIG biomarker may therefore be

considered as useful to identify patients at high and low risk of ARE. Moreover, the positive and negative post-test probability ratios at 40 days were 2.45 and 0.01, respectively. Patients who tested positive for MIG had a 2.45 fold greater risk of experiencing an ARE before the 40-day post-transplant time-point than after. In contrast, patients who tested negative for MIG had a 100 fold greater risk of experiencing an ARE after the 40-day time-point than before. These results also illustrate the better capacity of the MIG-based test to identify patients free of ARE during the first 40 days, but it should be not considered for populations with a different cumulative incidence of ARE.

Prognosis of a distant metastasis after breast cancer

In the article by Mook et al. [18], a 70-gene MammaPrint signature was studied to predict distant metastasis after surgery in patients with breast cancer. The HR group was made up of 439 patients with a positive 70-gene MammaPrint signature. The LR group comprised 525 patients presenting a negative 70-gene MammaPrint signature. The survival probabilities were significantly different between the HR and LR groups (hazard ratio=2.70, $p<0.001$). The distant metastasis survival probability up to 5 years was 0.80 in the HR group and 0.95 in the LR group. The cumulative probability of distant metastasis was 12% at 5 years. The 70-gene MammaPrint signature thus seemed to be a relatively sensitive but not very specific predictor of distant metastasis after surgery. Indeed, the sensitivity at 0.77 may be reasonable to control the number of false negatives for a prognosis up to 5 years, but the cost is a low specificity of 0.59 (Table 1). The sensitivity was even lower for a longer prognosis up to 10 years.

For a prognosis up to 5 years, the PPV and NPV were estimated at 0.20 and 0.95, respectively. Based on the 70-gene MammaPrint signature, patients considered in the LR

group had only a 5% risk of distant metastasis up to 5 years. In contrast, the accuracy of a positive test was small, with 80% of HR-classified patients without distant metastasis before 5 years. Mook et al. selected a population with a tumor size inferior to 2 cm for which the risk of a distant metastasis at 5 years was low. In France, the overall cumulative probability of distant metastasis at 5 years is lower (around 4.5%) in the same population (data from the BERENIS cohort, Nantes Institut de Cancérologie de l'Ouest). In this type of population, the PPV and NPV would be 8% and 98%, respectively (Figure 3). The NPV is thus very high, but the PPV is very low. As a result, 92% of patients with a positive MammaPrint signature would not have a distant metastasis.

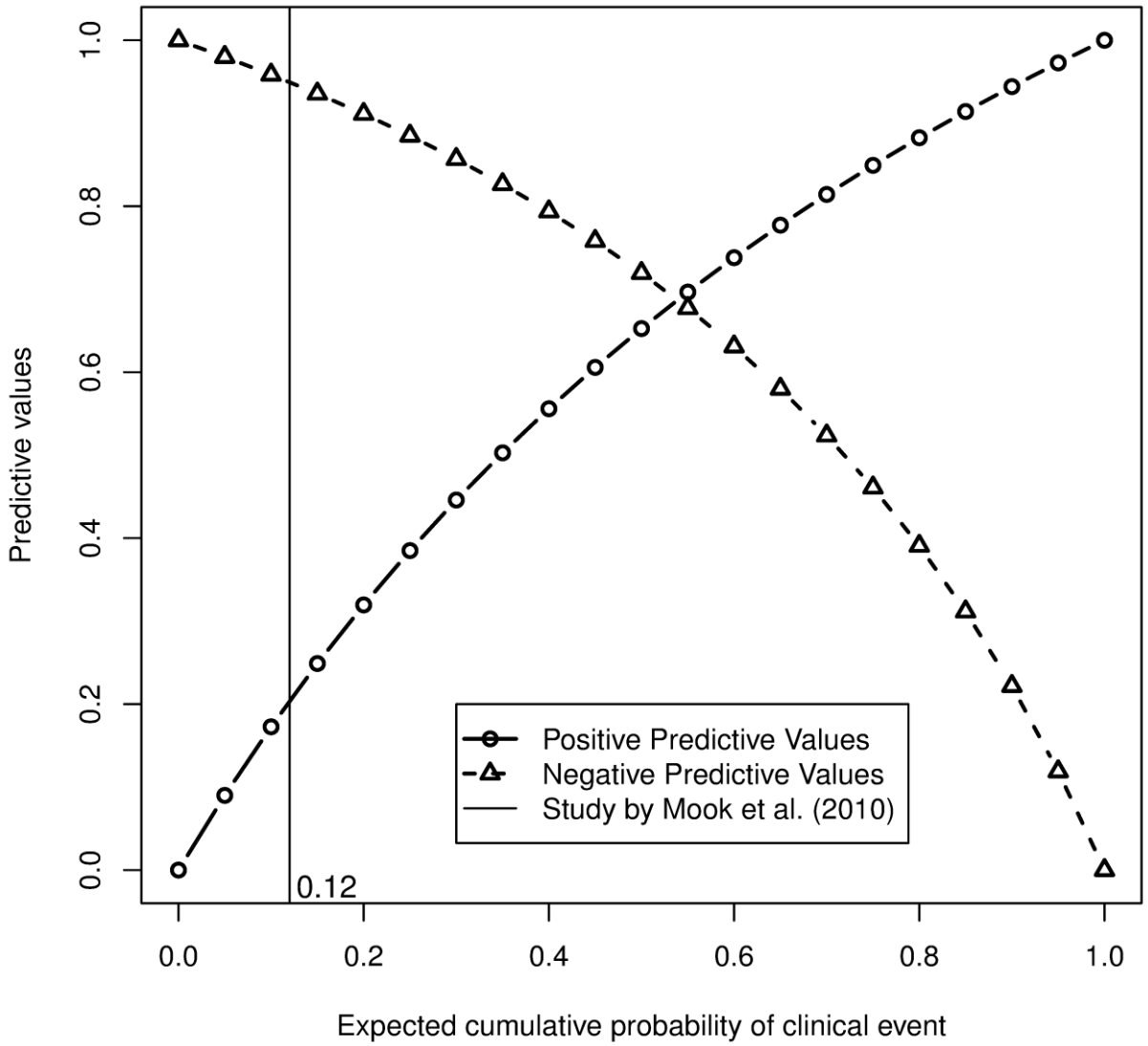


Figure 3. Positive and Negative Predictive Values regarding the Expected Cumulative Probability of Distant Metastasis up to 5 years post-surgery in Women with Breast Cancer. In the study by Mook et al. (2010), the Probability of Distant Metastasis before 5 years was 12%, the Positive and Negative Predictive Values were 20% and 95%, respectively.

The positive and negative likelihood ratios, estimated at 1.86 and 0.39, respectively, were very different from the expected values. Moreover, the positive and negative post-test probability ratios at 5 years were 0.25 and 0.05, respectively. Patients with a positive 70-

gene Mammaprint signature had a 0.25 fold greater risk of declaring a distant metastasis before 5 years post-treatment than after. Patients with a negative 70-gene Mammaprint signature had a 20 fold greater risk of declaring a metastasis after 5 years than before. In this population, the 70-gene MammaPrint signature may help to accurately identify patients without distant metastasis at 5 years post-surgery, but it would fail to identify patients with a future distant metastasis.

DISCUSSION

Many papers evaluating the prognostic capacity of markers use the same methodology: Kaplan-Meier survival curves, Log-Rank tests and/or the Cox model, while these methods are not adequate and should not be used in routine practice because they can lead to misinterpretation of results. Here, we have proposed simple equations to calculate time-dependent sensitivity, specificity, predictive values, likelihood ratios and post-test probability ratios from already published survival curves to confirm the prognostic capacity of a marker. We illustrate this methodology using two different examples. This work may put published results and recommendations into perspective by providing useful additional information. Nevertheless, it is always preferable for authors to systematically use the suitable methodology in the first place, to prove the prognostic capacity of a marker, in which case our approach would not be necessary. Time-dependent ROC curve [9, 19] can be directly estimated from individual data to estimate the prognostic capacity of a continuous marker. From the original definition of ‘C statistic’ proposed by Harrell et al. [20], different concordance indices have been developed and presented as measures of discrimination between patients with longer event-free survival and those with shorter event-free survival [21-24]. The net reclassification improvement has been recently developed and may be used to assess the performance of new prognostic markers [25].

Our approach is useful only *a posteriori* when the authors do not use appropriate methodologies.

Time-dependent sensitivity or specificity can strengthen the conclusions of papers reporting on the prognostic capacity of a marker. Markers that are significantly associated with the risk of the event may not necessarily be able to discriminate LR and HR subjects: the sensitivity may be acceptable at the cost of a low specificity, or vice versa. Time-dependent likelihood ratios provide complementary information. One advantage of sensitivity, specificity and likelihood ratios is their independence from the probability of the event. These indicators are therefore robust when applied to datasets from different countries and can even be used if the sample is not representative of the target population.

Moreover, time-dependent predictive values are relevant to clinical interpretations as they provide a direct probability of the event occurring or not before the prognostic time-point in a specific group. Time-dependent post-test probability ratios complete the prognostic marker evaluation by quantifying how many times higher the risk that the event occurs before the prognostic time-point is than that of the event occurring after the prognostic time-point. In contrast to the intrinsic indicators of the marker's prognostic capacity, predictive values and post-test probability ratios are dependent on the disease in probability, as illustrated in the results section. Major therapeutic decisions can be wrongly made if the cumulative probability of the event differs to that on which the marker was developed. We illustrated this approach using two applications.

Hauser et al. studied the urine biomarker MIG and the prediction of ARE early after kidney transplantation [17]. A negative MIG test indicated, with a high degree of confidence, that a patient would not suffer an ARE regardless of the cumulative probability of an ARE. Moreover, the negative post-test probability was very high: patients with a negative MIG test had a 100 times higher risk of suffering an ARE after the 40-day time-

point than before. In contrast, even though the risk of declaring an ARE before the 40-day time-point was higher than that after the 40-day time-point for patients with a positive MIG test, the positive predictive value was indicative of an unacceptably high false positive rate. If this test were to be used in a population with a more representative cumulative probability of ARE of around 9%, only 50% of MIG-positive patients would present an ARE within the 40-day period. If the strategy is to increase the immunosuppressive therapy for MIG-positive patients, this could lead to excessive treatment in a high percentage of patients. In a population with a 30% risk of ARE within the first 40 days, the expected PPV would be nearly 80% (Figure 2). Even with this relatively small error rate (20%), this biomarker would not be sufficiently accurate to blindly treat patients without any histological proof of the ARE. Thus, whereas this marker could be of interest to monitor patients routinely after kidney transplantation, to avoid unnecessary kidney biopsies in patients with a negative MIG test, the marker would be of limited use to increase immunosuppression in at-risk patients.

In the second paper, Mook et al. studied a population of women with a low risk of distant metastasis, i.e. a breast cancer tumor inferior to 2 cm in size and mostly treated by adjuvant therapy [18]. The 70-gene Mammaprint signature seemed to allow an acute selection of LR patients with a negative predictive value of 95% for a prognosis up to 5 years. Patient negative for the 70-gene Mammaprint signature had a 20 times greater risk of declaring a metastasis after 5 years than before. In contrast, if a supplementary chemotherapy is proposed for a positive signature, the PPV at 5 years indicates that 80% of HR patients may be over-treated. The positive post-test probability was 0.25, implying that 70-gene Mammaprint signature-positive patients had a 4 times greater risk of declaring a metastasis after 5 years than before. In normal circumstances, a marker with this type of positive post-test probability would not be considered for prognosis. Nevertheless, in the context of breast cancer, the overtreatment problem remains

questionable as the main risk concerns metastatic relapses which invariably lead to death. Actually, a women's choice for chemotherapy is based on the hypothesis of a survival gain of several months, with a known risk of overtreatment, even those for whom medical oncologist would not propose chemotherapy [26, 27]. Women generally prefer to receive unnecessary chemotherapy than not to be considered as HR whereas they should be. In view of the event in this context being terminal in nature, the optimal strategy tends to favor a high NPV, i.e. a small rate of women considered as LR whereas they should not be considered as such. Figure 3 shows that up to a cumulative probability of distant metastases of 22%, less than 10% of women are wrongly consider as LR. This indicates that the Mammaprint signature may not be used in populations with a higher risk of distant metastasis, as the number of women incorrectly considered as LR may be considered too high (greater than 10%).

The confidence intervals of the different time-dependent indicators would have been important to complete the description of the marker's prognostic capacity. However, only the variability of time-dependent predictive values may be easily assessed from the published confidence intervals of survival probabilities. The computation of the confidence intervals of the other indicators requires information about the marker and survival variability, which are not available from published survival curves.

As with many observational cohorts, survival curves may be misleading due to confounding factors of the prognostic marker. To fill in this major drawback, several authors recommend the drawing of adjusted survival curves, obtained for instance by an inverse probability weight (IPW) approach [28-30]. Even if this type of graphic representation should be used, most of survival curves presented in clinical papers are estimated from the Kaplan-Meier unadjusted estimator. An interesting feature of our approach is that it is still available for papers presenting adjusted survival curves by using

the survival estimations in in $SHR(t)$ and $SLR(t)$. Respecting the conclusions by Cole and Hernán [28], the corresponding adjusted time-dependent sensitivity will represent the sensitivity as if the patients with a positive test (classified in the HR group) have the same characteristics as the all sample. In parallel, the adjusted time-dependent specificity will represent the specificity as if the patients with a negative test (classified in the LR group) have the same characteristics as the all sample. This ensures no confounding factor between both groups. In extension, the other proposed indicators as time-dependent predictive values and post-test probability ratios can also be interpreted independently of confounding factors. We thus encourage the use of our methodology for such a type adjusted survival curve. Nevertheless, we did not find a paper based on IPW approach and devoted to the study of the capacity of a prognostic marker. In practice, for taking into account confounding factors, the authors always used Cox multivariate regression and the linear predictor as a prognostic composite score [31, 32]. Again, our proposed methodology can also be used based on survival curves stratified on such scoring system.

In conclusion, the methods we propose here can help to further interpret published results from papers reporting on raw or adjusted patient survival. Time-dependent sensitivity, specificity and likelihood ratios evaluate a marker's intrinsic prognostic capacity, while time-dependent predictive values and post-test probability ratios need to be interpreted according to the targeted population. Of course, any conclusion has to be additionally considered against the pathology being studied and the medical purpose of the prognostic marker.

ACKNOWLEDGEMENTS

This work was supported by a grant from the French National Research Agency ANR-11-JSV1-0008-01.

REFERENCES

1. *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.* Clin Pharmacol Ther, 2001. **69**(3): p. 89-95.
2. Buyse, M., et al., *Biomarkers and surrogate end points--the challenge of statistical validation.* Nat Rev Clin Oncol, 2010. **7**(6): p. 309-17.
3. Rector, T.S., et al., *Systematic Review of Prognostic Tests*, in *Methods Guide for Medical Test Reviews*. 2012, Agency for Healthcare Research and Quality (US): Rockville (MD).
4. Lachenbruch, P.A., et al., *Biomarkers and surrogate endpoints in renal transplantation: present status and considerations for clinical trial design.* American Journal of Transplantation, 2004. **4**(4): p. 451-7.
5. Pepe, M.S., et al., *Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker.* Am J Epidemiol, 2004. **159**(9): p. 882-90.
6. Ware, J.H., *The limitations of risk factors as prognostic tools.* N Engl J Med, 2006. **355**(25): p. 2615-7.
7. Spruance, S.L., et al., *Hazard Ratio in Clinical Trials.* Antimicrobial Agents and Chemotherapy, 2004. **48**(8): p. 2787-2792.
8. Foucher, Y., et al., *Prognostic markers: data misinterpretation often leads to overoptimistic conclusions.* Am J Transplant, 2012. **12**(4): p. 1060-1.
9. Heagerty, P.J., T. Lumley, and M.S. Pepe, *Time-dependent ROC curves for censored survival data and a diagnostic marker.* Biometrics, 2000. **56**(2): p. 337-44.
10. Carobbio, A., et al., *Leukocytosis and risk stratification assessment in essential thrombocythemia.* J Clin Oncol, 2008. **26**(16): p. 2732-6.

11. Kaplan, B., J. Schold, and H.U. Meier-Kriesche, *Poor predictive value of serum creatinine for renal allograft loss.* American Journal of Transplantation, 2003. **3**(12): p. 1560-5.

12. Wang, T.J., et al., *Multiple biomarkers for the prediction of first major cardiovascular events and death.* N Engl J Med, 2006. **355**(25): p. 2631-9.

13. Riley, R.D., et al., *Prognosis research: toward evidence-based results and a Cochrane methods group.* J Clin Epidemiol, 2007. **60**(8): p. 863-5; author reply 865-6.

14. Simel, D.L., J. Easter, and G. Tomlinson, *Likelihood ratios, sensitivity, and specificity values can be back-calculated when the odds ratios are known.* J Clin Epidemiol, 2012.

15. Deeks, J.J. and D.G. Altman, *Diagnostic tests 4: likelihood ratios.* BMJ, 2004. **329**(7458): p. 168-9.

16. Jaeschke, R., G. Guyatt, and D. Sackett, *Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test: B. What Are the Results and Will They Help Me In Caring for My Patients?* JAMA 1994. **271**: p. 703-7.

17. Hauser, I.A., et al., *Prediction of acute renal allograft rejection by urinary monokine induced by IFN-gamma (MIG).* J Am Soc Nephrol, 2005. **16**(6): p. 1849-58.

18. Mook, S., et al., *Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature.* Ann Surg Oncol, 2010. **17**(5): p. 1406-13.

19. Heagerty, P.J. and Y. Zheng, *Survival model predictive accuracy and ROC curves.* Biometrics, 2005. **61**(1): p. 92-105.

20. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in medicine, 1996. **15**(4): p. 361-387.

21. Chambless, L.E. and G. Diao, *Estimation of time-dependent area under the ROC curve for long-term risk prediction.* Statistics in medicine, 2006. **25**(20): p. 3474-3486.

22. Pencina, M.J. and R.B. D'Agostino, *Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation.* Statistics in medicine, 2004. **23**(13): p. 2109-2123.

23. Pencina, M.J., R.B. D'Agostino, Sr., and L. Song, *Quantifying discrimination of Framingham risk functions with different survival C statistics.* Statistics in medicine, 2012. **31**(15): p. 1543-1553.

24. Uno, H., et al., *On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data.* Statistics in medicine, 2011. **30**(10): p. 1105-1117.

25. Pencina, M.J., et al., *Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond.* Statistics in medicine, 2008. **27**(2): p. 157-172; discussion 207-212.

26. Mandelblatt, J.S., et al., *Breast cancer adjuvant chemotherapy decisions in older women: the role of patient preference and interactions with physicians.* J Clin Oncol, 2010. **28**(19): p. 3146-53.

27. Moumjid, N., et al., *Clinical issues in shared decision-making applied to breast cancer.* Health Expect, 2003. **6**(3): p. 222-7.

28. Cole, S.R. and M.A. Hernan, *Adjusted survival curves with inverse probability weights.* Comput Methods Programs Biomed, 2004. **75**(1): p. 45-9.

29. Westreich, D., et al., *Time scale and adjusted survival curves for marginal structural cox models.* Am J Epidemiol, 2010. **171**(6): p. 691-700.

30. Xie, J. and C. Liu, *Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data.* Stat Med, 2005. **24**(20): p. 3089-110.

31. Foucher, Y., et al., *A clinical scoring system highly predictive of long-term kidney graft survival.* Kidney International, 2010. **78**(12): p. 1288-94.

32. Spiegelhalter, D.J., *Probabilistic prediction in patient management and clinical trials.* Stat Med, 1986. **5**(5): p. 421-33.

Appendix

The failure status at any time t is noted as the following counting process $D(t) = 1$ if the failure is before time t and $D(t) = 0$ if the failure is not before time t. NHR represents the number of High Risk patients and NLR represents the number of Low Risk patients. $SHR(t)$ and $SLR(t)$ represent the survival probabilities at time t of these high and low risk patients respectively. The sensitivity and specificity can be calculated as follows:

$$Se(t) = P(HR|D(t) = 1) = \frac{P(HR, D(t) = 1)}{P(D(t) = 1)} = \frac{P(D(t) = 1|HR) \times P(HR)}{P(D(t) = 1, HR) + P(D(t) = 1, LR)}$$

$$= \frac{P(D(t) = 1|HR) \times P(HR)}{P(D(t) = 1|HR) \times P(HR) + P(D(t) = 1|LR) \times P(LR)}$$

$$= \frac{\big(1 - P(D(t) = 0|HR)\big) \times P(HR)}{\big(1 - P(D(t) = 0|HR)\big) \times P(HR) + \big(1 - P(D(t) = 0|LR)\big) \times P(LR)}$$

$$= \frac{\big(1 - S(t|HR)\big) \times P(HR)}{\big(1 - S(t|HR)\big) \times P(HR) + \big(1 - S(t|LR)\big) \times P(LR)}$$

$$= \frac{(1 - SHR) \times NHR}{(1 - SHR) \times NHR + (1 - SLR) \times NLR}$$

$$Sp(t) = P(LR|D(t) = 0) = \frac{P(LR, D(t) = 0)}{P(D(t) = 0)} = \frac{P(D(t) = 0|LR) \times P(LR)}{P(D(t) = 0, HR) + P(D(t) = 0, LR)}$$

$$= \frac{P(D(t) = 0|LR) \times P(LR)}{P(D(t) = 0|HR) \times P(HR) + P(D(t) = 0|LR) \times P(LR)}$$

$$= \frac{S(t|LR) \times P(LR)}{S(t|HR) \times P(HR) + S(t|LR) \times P(LR)}$$

$$= \frac{SLR(t) \times NLR}{SHR(t) \times NHR + SLR(t) \times NLR}$$

Using the Bayes theorem, the positive and negative predictive values can be calculated as follows:

$$PPV(t) = P(D(t) = 1|HR) = \frac{P(D(t) = 1, HR)}{P(HR)} = \frac{P(HR|D(t) = 1) \times P(D(t) = 1)}{P(D(t) = 1, HR) + P(D(t) = 0, HR)}$$

$$= \frac{P(HR|D(t) = 1) \times P(D(t) = 1)}{P(HR|D(t) = 1) \times P(D(t) = 1) + P(HR|D(t) = 0) \times P(D(t) = 0)}$$

$$= \frac{Se(t) \times P(D(t) = 1)}{Se(t) \times P(D(t) = 1) + \big(1 - P(LR|D(t) = 0)\big) \times \big(1 - P(D(t) = 1)\big)}$$

$$= \frac{Se(t) \times P(D(t) = 1)}{Se(t) \times P(D(t) = 1) + (1 - Sp(t)) \times \big(1 - P(D(t) = 1)\big)}$$

$$NPV(t) = P(D(t) = 0|LR) = \frac{P(D(t) = 0, LR)}{P(LR)} = \frac{P(LR|D(t) = 0) \times P(D(t) = 0)}{P(D(t) = 0, LR) + P(D(t) = 1, LR)}$$

$$= \frac{Sp(t) \times (1 - P(D(t) = 1))}{P(LR|D(t) = 0) \times P(D(t) = 0) + P(LR|D(t) = 1) \times P(D(t) = 1)}$$

$$= \frac{Sp(t) \times (1 - P(D(t) = 1))}{Sp(t) \times \big(1 - P(D(t) = 1)\big) + \big(1 - P(HR|D(t) = 1)\big) \times P(D(t) = 1)}$$

$$= \frac{Sp(t) \times (1 - P(D(t) = 1))}{Sp(t) \times \big(1 - P(D(t) = 1)\big) + (1 - Se(t)) \times P(D(t) = 1)}$$

If disease probability is not given, it can also be recalculated from the survival curves as follows:

$$
\begin{aligned}
P(D(t) = 1) &= P(D(t) = 1|HR) \times P(HR) + P(D(t) = 1|LR) \times P(LR) \\
&= \big(1 - P(D(t) = 0|HR)\big) \times P(HR) + \big(1 - P(D(t) = 0|LR)\big) \times P(LR) \\
&= \frac{\big(1 - S(t|HR)\big) \times NHR + \big(1 - S(t|LR)\big) \times NLR}{NHR + NLR} \\
&= \frac{NHR - S(t|HR) \times NHR + NLR - S(t|LR) \times NLR}{NHR + NLR} \\
&= 1 - \frac{SHR(t) \times NHR + SLR(t) \times NLR}{NHR + NLR}
\end{aligned}
$$

Time-dependent likelihood ratios allow the following relationship between pre-test probabilities and post-test probability ratios, noted $PT^+(t)$ and $PT^-(t)$:

$$
\begin{aligned}
PT^+(t) &= \frac{P(D(t) = 1|HR)}{P(D(t) = 0|HR)} \\
&= \frac{P(D(t) = 1)}{P(D(t) = 0)} \times \frac{P(D(t) = 1|HR)P(HR)}{P(D(t) = 1)} \times \frac{P(D(t) = 0)}{P(D(t) = 0|HR)P(HR)} \\
&= \frac{P(D(t) = 1)}{P(D(t) = 0)} \times \frac{P(HR|D(t) = 1)}{P(HR|D(t) = 0)} = \frac{P(D(t) = 1)}{P(D(t) = 0)} LikR^+(t)
\end{aligned}
$$

$$
\begin{aligned}
PT^-(t) &= \frac{P(D(t) = 1|LR)}{P(D(t) = 0|LR)} \\
&= \frac{P(D(t) = 1)}{P(D(t) = 0)} \times \frac{P(D(t) = 1|LR)P(LR)}{P(D(t) = 1)} \times \frac{P(D(t) = 0)}{P(D(t) = 0|LR)P(LR)} \\
&= \frac{P(D(t) = 1)}{P(D(t) = 0)} \times \frac{P(LR|D(t) = 1)}{P(LR|D(t) = 0)} = \frac{P(D(t) = 1)}{P(D(t) = 0)} LikR^-(t)
\end{aligned}
$$