Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A balanced hazard ratio for risk group evaluation from survival data

Samuel Branders [a] and Pierre Dupont [a]*

Common clinical studies assess the quality of prognostic factors, such as gene expression signatures, clinical variables or environmental factors, and cluster patients into various risk groups. Typical examples include cancer clinical trials where patients are clustered into high or low risk groups. Whenever applied to survival data analysis, such groups are intended to represent patients with similar survival odds and to select the most appropriate therapy accordingly. The relevance of such risk groups, and of the related prognostic factors, is typically assessed through the computation of a hazard ratio. We first stress three limitations of assessing risk groups through the hazard ratio: 1) it may promote the definition of arbitrarily unbalanced risk groups, 2) an apparently optimal group hazard ratio can be largely inconsistent with the p-value commonly associated to it, 3) some marginal changes between risk group proportions may lead to highly different hazard ratio values. Those issues could lead to inappropriate comparisons between various prognostic factors. Next we propose the balanced hazard ratio to solve those issues. This new performance metric keeps an intuitive interpretation and is as simple to compute. We also show how the balanced hazard ratio leads to a natural cut-off choice to define risk groups from continuous risk scores. The proposed methodology is validated through controlled experiments for which a prescribed cut-off value is defined by design. Further results are also reported on several cancer prognosis studies and the proposed methodology could be applied more generally to assess the quality of any prognostic markers.

Keywords: Survival data; hazard ratio; risk groups; prognostic factors

1. Introduction

Several clinical studies monitor the survival of patients while assessing new prognostic markers, such as gene expression signatures, clinical variables or environmental factors. Cancer clinical trials are typical examples of such studies where the time to an event, *e.g.* metastasis, relapse or death, is recorded. Such data is also usually censored whenever a patient disappears, possibly by dying from another cause, or simply leaves the study before the event of interest occurs. Prognostic indexes made of one or several bio-markers are analyzed in such studies to improve the treatment guidelines. For instance,

[a] *Université catholique de Louvain – ICTEAM/Machine Learning Group, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium*
* *Correspondence to: Université catholique de Louvain – ICTEAM/Machine Learning Group, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium. E-mail: pierre.dupont@uclouvain.*

Contract/grant sponsor: F.R.S. - FNRS - Télévie (Grant number FC 88088).

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

one decides according to a patient estimated risk whether he needs adjuvant chemotherapy. To perform this task, prognostic indexes commonly split the patients into a few risk groups [1, 2, 3, 4, 5]. Such risk groups aim at clustering patients with a similar prognosis and guide the associated treatments.

From a data analysis viewpoint, the very definition of groups may be simple when they are specifically associated to a given factor of interest. For example, one may analyze the survival times between smokers and non-smokers to assess whether such factor plays an important role. In a more complex yet common setting, risk groups are derived from continuous risk scores. Indeed, these scores aggregate the influence of several potential co-factors, and the risk groups are obtained by defining cut-off values (*i.e.* thresholds) on those scores. In such a situation, the definition of risk groups requires the choice of the factors to be considered, a methodology to combine the numeric values from each factor into a global score, also known as prognostic index, and the setting of thresholds on those scores. In other words, the definition of risk groups becomes an unsupervised clustering problem for which one typically aims at associating patients with similar survival times to the same risk group while maximizing the differences between groups. In particular, one expects the patients among the highest risk group(s) to experience the event (*e.g.* metastasis or relapse) before patients from other groups.

The risk scores and the associated risk groups are commonly assessed through the computation of a hazard ratio [6] or similar performance metrics, such as the concordance index [7]. Our first contribution is to argue why these metrics, while being perfectly sound to assess risk *scores*, are much less appropriate to evaluate risk *groups*. Consequently, these metrics are also problematic to assess the quality of some co-factors of interest as prognostic markers to distinguish between risk groups.

Part of the issue comes from the choice of the threshold (or cut-off) values on risk scores to obtain the risk groups. Indeed, these cut-offs often look arbitrarily chosen or fixed on a given collection of samples without a proper assessment of the generalization capability of such choice on new and independent samples [8]. More fundamentally, such an assessment requires an appropriate performance metric and we stress the inadequacy of the popular hazard ratio (HR), as well as common alternatives, for this task. We show in particular that the HR may be artificially increased by considering highly unbalanced groups: an extremely unbalanced choice would, for instance, consider a single patient with the shortest survival time (or the highest risk score) as the unique member of the high risk group, while putting all other samples in a presumably low risk group. Such an extreme choice is likely to lead to a very high HR but is unlikely to be valuable from an accurate prognosis viewpoint. A perfect balance, say 50%/50% between high and low risk groups, needs not be relevant either. Unless some prior information exists about the relative size of the risk groups, which is rarely the case when assessing the prognostic values of new candidate markers, the definition of risk groups looks ill-defined when assessed through the standard HR. We stress that this problem occurs beyond extremely unbalanced cases as a consequence of the HR measure exhibiting potentially many local optima and being particularly non-smooth. Therefore, very small modifications of the proportions between risk groups (through marginal cut-off modifications) can lead to highly different HR values. We also show that maximizing HR versus minimizing its associated p-value could lead to drastically different risk groups.

Our second contribution is the definition of a new performance metric, called the *balanced hazard ratio* (BHR) to fix the issues raised above. The BHR keeps an intuitive interpretation and is as simple to compute as the original HR, meaning that it can be easily used by clinician accustomed to the hazard ratio. Yet, the BHR penalizes artificially unbalanced risk groups and, more generally, offers a smoother profile with a natural optimum. Interestingly, such optimum is data-dependent and needs not correspond to a perfect balance between groups. Our third contribution is to show that the BHR leads to a natural definition of cut-off values on risk scores to define associated risk groups.

We illustrate the proposed methodology on breast cancer studies assessing the quality of prognostic gene signatures. Section 2 briefly describes those studies, which are chosen here as running examples. Yet we believe that our conclusions fully apply to the general evaluation of risk groups from survival data. Further results with other prognosis models and additional cancer studies (colon, ovarian) are reported in the supplementary materials. They are fully consistent with those described here. Section 3 describes the definition of risk groups from cut-off values on risk scores provided by common

prognostic indexes. It illustrates, on some running examples, the influence of such cut-offs on the evaluation of risk groups. This initial analysis calls for dedicated evaluation metrics. Section 4 precisely describes the original hazard ratio and discusses why its use is problematic to assess risk groups. Alternative performance metrics proposed in the literature are described in section 5. We compare them to the hazard ratio and we discuss their relevance for evaluating risk groups. Section 6 presents the balanced hazard ratio. We argue why it is well designed for evaluating risk groups, while keeping a natural interpretation along the same lines as the original hazard ratio. Section 7 describes how the balanced hazard ratio, originally defined for two risk groups, can be easily generalized to an arbitrary number of risk groups. Section 8 describes how the BHR can be used to choose natural cut-off values on risk scores. Section 9 further illustrates the soundness of the proposed methodology on controlled experiments for which an underlying threshold between risk groups is fixed by design. We conclude our work and discuss additional perspectives in section 10.

2. Illustrative clinical studies

Risk group prediction and evaluation are illustrated in this work on known clinical studies of breast cancer prognosis. These studies offer a variety of data sets publicly available from the GEO database with existing prognostic markers, often made of gene expression signatures. The high prevalence of breast cancer and the availability of several prognostic indexes for essentially the same task drove our choice on those examples but our conclusions are aimed to be applicable to any evaluation of risk groups from prognostic indexes and survival data. In particular, the supplementary materials present similar results with additional prognosis models for breast cancer, and further results on colon and ovarian cancers.

We consider in particular the so-called Gene76 and Gene70 gene expression signatures, related to MammaPrint [5, 3]. MammaPrint is a test sold to assess the risk of metastasis in breast cancer. This test was build from the Gene70 model and the signature described in [5].

The Gene76 prognostic model is build from 76 genes to identify patients who developed distant metastasis within 5 years. All patients considered are node negative and untreated. The gene expression and survival data for this study form the Veridex (VDX) data set [3]: $n = 344$ samples, GEO accession numbers GSE2034 and GSE5327.

The Gene76 prognostic model has been further validated on an independent study conducted by the TRANSBIG (TBG) consortium. The TBG data set includes untreated patients with primary breast cancer and a node negative status [9]: $n = 198$ samples, GEO accession number GSE7390.

The data set UNT comes from a study investigating the links between histopathological grades and gene expressions [10]. To focus on comparable data sets, we consider only untreated patients from this study with a node negative status after removing samples also present in VDX or TBG: $n = 84$, GEO accession number GSE2990.

In all those studies, distant metastasis is used as end point and no information is available about possible competing risks. Gene expression data are measured on the Affymetrix HGU133a microarray platform. All data sets were summarized according to the MAS5.0 procedure and represented in $\log_2$ scale. Practical experiments were conducted using the R statistical language, including specific breast cancer prognostic models implemented in the `genefu` R package [11] from Bioconductor.

3. From risk scores to risk groups

Throughout this paper, we use the following notations. Given some survival data from n patients, t_i denotes the time at which patient i experienced the event (for example, a distant metastasis) or the time at which he was censored. The binary variable δ_i takes value 1 if the patient experienced the event, 0 otherwise. For each patient i, we consider a risk score r_i and a risk group g_i as detailed below.

A common prognostic score is defined by the Nottingham prognostic index (NPI) for primary breast cancer [1]. It is formally defined as follows:

$$NPI = S * 0.2 + G + N \tag{1}$$

where

- S is the tumor size in centimeters,
- G represents the tumor grade from 1 to 3 (good, moderate or poor),
- N has a value of 1, 2, or 3 according to the number of lymph nodes involved: 0, 1 to 3, or more than 3, respectively.

The NPI defines a continuous score from 3 histo-pathological predictors S, G and N. A higher score corresponds to a patient predicted to be at higher risk. Such index is commonly associated to cut-off values on the scores to cluster patients into different risk groups:

- Good prognosis : $NPI < 3.4$
- Moderate prognosis: $3.4 \leq NPI < 5.4$
- Poor prognosis : $NPI \geq 5.4$

The NPI is a simple linear model aggregating the prognostic influence of several covariates. More generally, such model defines a risk score r_i, for the patient i, as $r_i = \boldsymbol{w}^\top \boldsymbol{x}$, where $\boldsymbol{x}$ is the vector of covariates (or predictors) and $\boldsymbol{w}$ a weight vector[†]. The same formula applies to alternative prognostic indexes for which $\boldsymbol{x}$ is made of gene expression values forming a gene signature, possibly incorporated in a generalized linear model such as a logistic regression or a Cox proportional hazards model [6].

High versus low risk groups of patients directly follow from the definition of a cut-off θ on the risk scores. Specifically, the risk score r_i of patient i is continuous while his risk group is defined through the binary variable g_i either as 0 (low risk) or 1 (high risk).

$$g_i = \begin{cases} 0 & \text{if} \quad r_i < \theta \\ 1 & \text{if} \quad r_i \geq \theta \end{cases} \tag{2}$$

We note that our first example, NPI, actually defines 2 cut-offs and 3 risk groups but, for the sake of our analysis, one can focus initially on the first cut-off value (3.4) to distinguish between low risk (good prognosis) and high risk (moderate or poor prognosis) patients. Generalization to more than 2 risk groups is discussed in section 7.

To illustrate the influence of the cut-off choice, we consider the Gene76 prognostic index on the VDX dataset. This index first partitions the patients depending on their estrogen receptor status, being either positive or negative. For each status, the risk score is defined through a specific combination of univariate Cox models. Figure 1 reports the Kaplan-Meier curves for the 2 risk groups as defined by this model and its original cut-off. The specific details on how this original cut-off was chosen are further discussed in section 8. We focus here on the impact of this choice versus possible alternatives. Kaplan-Meier curves present the proportion of patients still at risk (*i.e.* without having experienced the event) along the follow-up time (expressed here in years). Crosses on the curves represent censored data. There is one curve for each risk group with the corresponding number of patients in each group being reported below the x-axis. Informally speaking, the more separated the 2 curves, the better the prognostic index and its underlying predictors as prognostic markers. To evaluate such a difference, the hazard ratio (HR) is commonly considered, together with its 95% confidence interval and a p-value of a statistical test assessing whether HR significantly differs from 1.

The formal definition of the hazard ratio is detailed in section 4. At this point, one can simply state that the higher the HR the better. In this regard, Gene76 looks to be a good prognostic index, as evaluated on the VDX dataset, since its HR is

[†]In the particular case of NPI, $\boldsymbol{w}^\top = [0.2\ 1\ 1]$.

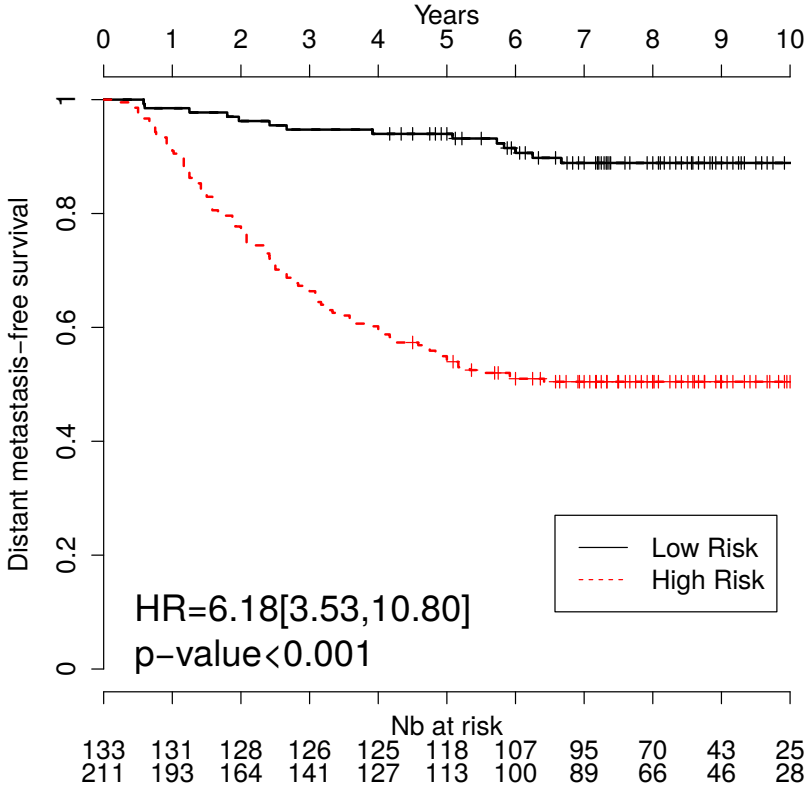
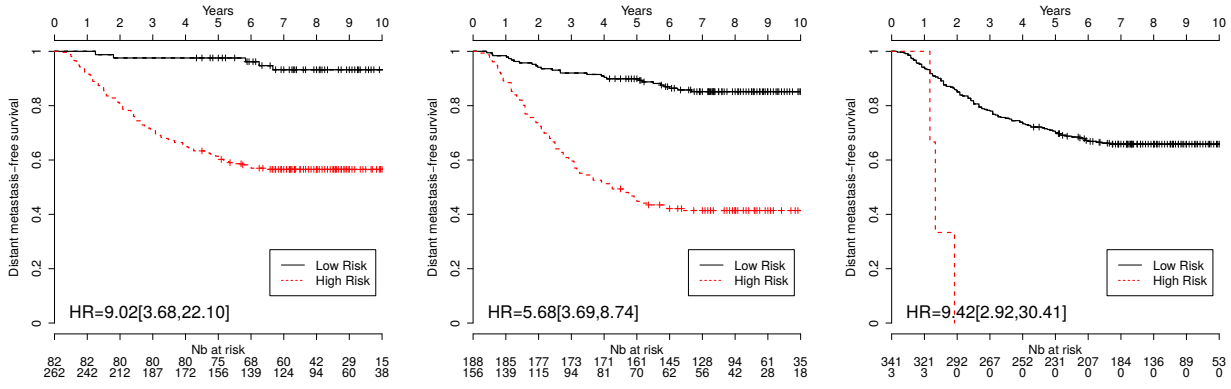


Figure 1. Risk groups on the VDX dataset with the original Gene76 cut-off.



(a) (b) (c)

Figure 2. Risk groups on the VDX data set defined by the Gene76 model with alternative cut-offs.

high (6.18), with a 95% confidence interval = $[3.53, 10.80]$ and a very small p-value. Yet, this would need to be confirmed on additional samples independently from those used to estimate this model, as further discussed in section 8.

We focus here on a different issue. Alternative cut-off choices are possible and could lead to strongly different HR evaluation. The figure 2 presents the survival curves of the high and low risk groups from the same Gene76 model evaluated on the same dataset (VDX) but with alternative cut-offs.

Changing the cut-offs logically affects the number of patients in the low versus high risk groups but it also largely modifies the HR values. In particular, a HR value larger than 9 is reported in figures 2(a) and 2(c), which is much higher than the original HR. In the first case, the number of patients in the low risk group has been decreased (from 133 to 82 at time 0). In the latter case, nearly all patients (341 out of 344 at time 0) now belong to the low risk group, which forms a

particularly imbalanced splitting between risk groups. Such observations tend to show that an appropriate cut-off choice is critical. Before revisiting this question, we argue in section 4 why assessing risk group prediction through a group hazard ratio is problematic.

The results presented in figure 2 will be used as running examples throughout this paper. Similar results are presented in the supplementary material with additional prognostic models of breast cancer and further results on colon and ovarian cancers.

4. Group Hazard Ratio

The group hazard ratio (HR) evaluates the difference between survival curves computed by a Cox proportional hazards model. It represents the increase in the risk of event between the low and high risk groups. When used to evaluate risk groups, the hazard ratio is computed with a Cox model using the binary group variable g_i as single covariate. The hazard function $h_i(t)$ for a patient i is then written as :

$$h_i(t) = h_0(t) \exp(\beta g_i) \tag{3}$$

Since g_i equals 0 for the patients in the low risk group, $h_0(t)$ is the hazard function for the patients of this group, while the multiplicative factor $\exp(\beta)$ applies to the high risk group.

$$h_i(t) = \begin{cases} h_0(t) & \text{if} \quad g_i = 0 \\ h_0(t) \exp(\beta) & \text{if} \quad g_i = 1 \end{cases} \tag{4}$$

Consequently, this multiplicative factor is precisely the hazard ratio between both groups: $HR = \exp(\beta)$. The higher the hazard ratio, the higher the difference between the two hazard functions and thus between the survival curves.

As illustrated in figure 2, a specific cut-off choice on the risk scores may largely influence the proportion of samples in each risk groups and the resulting HR values. Figure 3 generalizes this analysis by reporting the HR values (along the left y-axis) obtained for all possible cut-offs defining a proportion of samples in the low risk group varying from 0% to 100% (along the x-axis). Results presented in figure 2 correspond to 3 specific cut-offs, hence 3 specific points on this HR curve, illustrated by black, red and green dots respectively while the original Gene76 cut-off is represented with a dashed line. The HR value can clearly be artificially increased by considering extremely unbalanced risk groups, which would nevertheless be uninformative from a prognostic point of view. The problem is even more serious since this HR (plain) curve exhibits many local optima and is far from being smooth. In other words, marginal changes in the relative proportions between risk groups may drastically affect the observed hazard ratio while potentially modifying, positively or negatively, the estimated quality of the predictors used as prognostic markers. Finally, due to those fluctuations, the evolution of the HR values are inconsistent with the associated p-values (dashed orange line, $-\log$ scale along the right y-axis). Hence risk groups maximizing the HR value (*e.g.* the green dot) is largely different those minimizing the associated p-values (`e.g.` the red dot).

All the above results illustrate that using the hazard ratio to assess the quality of prognostic markers to discriminate between risk groups is highly questionable.

We stress that the issues raised here directly follow from the use of a *discrete* indicator variable g_i in the definition (4) of the group hazards. Another classical definition of HR considers a *continuous* risk score r_i instead of the discrete (and here binary) g_i. In the continuous case, let us assume for example $HR = 2$ while comparing the risk scores $r_i > r_j$, associated to patients i and j still at risk. Such a HR value would mean that the probability of experiencing the event is twice as large for patient i. This is a perfectly sound use of the HR measure to assess the relevance of risk scores. The problems raised above appear whenever cut-offs are chosen on those scores and discrete groups are defined accordingly. Such

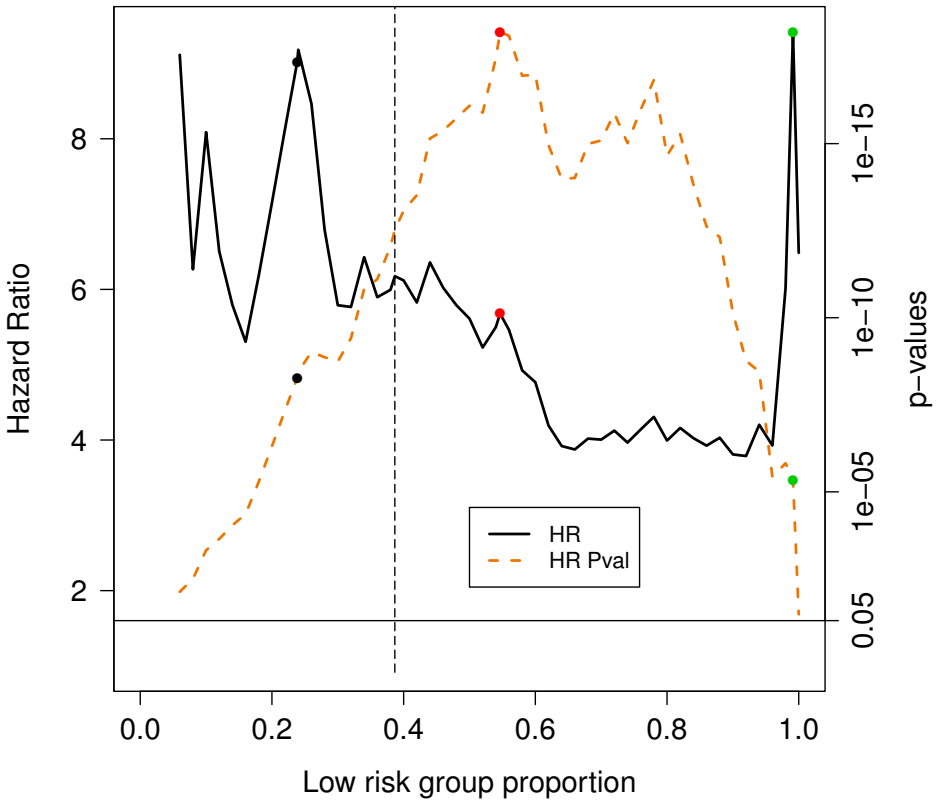


Figure 3. Plain line: hazard ratio of the Gene76 model on the VDX dataset while varying the proportions in each risk group through different cut-off choices (HR value on the left y-axis). The dashed vertical line corresponds to the original cut-off. Dashed orange line: associated p-values plotted in $-\log$ scale (right y-axis, the higher the better on such a plot).

discretization needs however to be considered as it is routinely used by clinicians to decide whether a specific treatment should be given to a patient. Such decision is indeed often based on the assignment of the patient to a particular risk group. We introduce in section 6 a novel performance metric, called the balanced hazard ratio, which fixes those issues while keeping an interpretation similar to the original hazard ratio. In the meantime, we revisit alternative performance metrics defined in the literature.

5. Alternative performance metrics

Several performance metrics have been proposed in the literature to evaluate risk group prediction models. We briefly review here some popular metrics including the concordance index, the logrank test, the SEP and the pair sensitivity and specificity.

5.1. Concordance index

The concordance index (C-index) measures to which extent the risk groups are concordant with the time to event, that is whether the patients in the high risk group actually experience the event before the patients in the low risk group [7].

The C-index specifically relies on the notion of comparable pairs. A pair of patients $\{i, j\}$ is comparable if patient i experiences the event while the patient j is still at risk (not censored and not having experienced the event) at time t_i. The

two patients i and j must also be in different risk groups. Such comparable pair is concordant if the patient i, experiencing the event earlier, belongs to the high risk group. The C-index lies between 0 and 1 as it estimates the probability for a comparable pair of patients to be concordant. This estimate is the number of concordant comparable pairs divided by the number of comparable pairs:

$$\text{C-Index} = \frac{\sum_{i,j} 1\{g_i > g_j\}}{|\{\ \{i,j\} \mid g_i \neq g_j,\ \delta_i = 1,\ t_i < t_j\}|} \tag{5}$$

5.2. *Logrank test*

The logrank is the statistics of a test to assess whether there is a significant survival difference between risk groups [12]. The null hypothesis of such a test assumes no survival difference across groups. The logrank measures differences during the follow-up in number of events between what is observed in a group and what is expected under the null hypothesis. A high logrank implies that there is higher than expected rate of events in the high risk group and more evidence against the null hypothesis.

The logrank statistics is a sum of hypergeometric random variables (one variable for each time step t) which approximately follows a normal distribution, from which a p-value can be easily computed.

5.3. *SEP*

The SEP metric compares the risk of each risk groups with the risk of the entire population. SEP is a weighted geometric mean of the absolute relative risks between the risk groups and the global population [13].

$$SEP = \exp\left(\sum_k \frac{n_k}{n} |\beta_k|\right) \tag{6}$$

where n_k is the number of patients in the k^{th} risk group, and β_k is the log hazard ratio between the k^{th} risk group and the whole population.

5.4. *Sensitivity and specificity*

The sensitivity/specificity is the proportion of positive/negative samples identified as such by a model. In the survival context, the sensitivity $SE(t)$ at time t is the proportion of positive patients, *i.e.* classified in the high risk group, among those patients experiencing the event before time t. The specificity $SP(t)$ at time t is the proportion of negative patients, *i.e.* classified in the low risk group, among those patients still at risk just after time t.

$$SE(t) = \frac{\sum_{i|t_i \leq t, \delta_i = 1} g_i}{\sum_{i|t_i \leq t, \delta_i = 1} 1} \quad SP(t) = \frac{\sum_{i|t_i > t} (1 - g_i)}{\sum_{i|t_i > t} 1} \tag{7}$$

The above measures are defined for a specific time index t as a consequence of the fact that risk groups evolve through time since some patients are censored and some experience the event. The sensitivity/specificity trade-off could thus be reported as a curve evolving through the time of follow up. A common alternative is to fix a specific time of interest. For instance, in the context of breast cancer prognosis, $t = 5$ years after treatment is commonly accepted as a critical value from a clinical viewpoint.

To summarize the specificity/sensitivity in a single measure, it is convenient to define the balanced classification rate (BCR) as the arithmetic average between both $BCR(t) = \frac{SE(t)+SP(t)}{2}$.

BCR values lies between 0 and 1. A perfect prognostic index would have a BCR of 1. Uniform random guessing between risk groups has an expected BCR of 0.5 while a lower BCR would correspond to an even worse prognosis (*e.g.* inverting risk groups). We note that, unlike classification accuracy, BCR is insensitive to the group relative proportions, hence the

name *balanced*. Unlike ROC analysis, BCR also trivially generalizes to more than 2 risk groups, simply as the arithmetic average between classification rates in each group.

5.5. Discussion

We study the relative behaviors of the various performance metrics on our running example. Figure 4 further extends figure 3 by reporting the four metrics under study after rescaling all of them between 0 and 1 to ease the comparison (BCR is reported at 5 years after treatment, for all possible risk score cut-offs).

We note that the C-index behaves very similarly to the hazard ratio: one can trivially optimize them while considering artificially unbalanced groups, they both exhibit many local optima and sharp fluctuations for marginally different group proportions. In contrast, the logrank test, the SEP and the BCR look more appropriate as they offer quite smoother curves with a similar global optimum observed for more balanced groups. We note that this optimum is data dependent and needs not correspond exactly to a 50%/50% balance between groups. Yet those measures are not fully satisfactory either. The logrank statistics is a sum of hypergeometric random variables which does not offer an easy interpretation in terms of survival times differences between risk groups, unlike the hazard ratio and the C-index. Yet, actual survival times are key features for the patients and for choosing appropriate treatments.

The SEP metric offers a more direct interpretation than the logrank and provides an estimate of the degree of separation of the different risk groups. Both metrics behave very similarly but also share a common disadvantage: they are insensitive to the ordering of the risk groups. In other words, those values are unchanged after inverting risk groups and wrongly assigning the patients with a lower risk score to the higher risk group. This problem becomes even more serious with more than 2 risk groups (see section 7).

As for the BCR, it is not a distance measure between survival curves but rather a balanced measure of classification rates in each group. It also depends on a critical time value (here chosen at 5 years after treatment) which is somewhat arbitrary and moreover largely dependent on the pathology under study. Finally, specificity/sensitivity measures are not perfectly suited to censored data because they simply ignore the patients who did not experience the event and have been censored before the critical time considered.

Section 6 introduces a novel performance metric to address those issues.

6. Balanced Hazard Ratio

The balanced hazard ratio (BHR) computes the hazard ratio between three curves: the survival curves of the high and low risk groups (as for HR) and a third global survival curve over all patients (we present a generalization to more than 2 risk groups in section 7). Each sample is now considered as a member of 2 groups: its actual risk group ($g_i = -1$, for low risk, or $g_i = 1$, for high risk) and the global risk group ($g_i = 0$) for all patients. Such a global risk group represents the hazard (or survival time) over the whole population of patients and one measures now how much each specific risk group departs from the global curve. Figure 5 illustrates those survival curves on our running example with the 3 different proportions between risk groups considered so far.

The hazard function is now defined over those 3 groups:

$$h_i(t) = h_0(t) \exp(\beta g_i), \text{with } g_i = -1, 0, \text{ or } 1 \tag{8}$$

The quantity $h_0(t)$ represents here the hazard of the whole population, $h_0(t)/\exp(\beta)$ the hazard of the low risk group and $h_0(t)\exp(\beta)$ the hazard of the high risk group. The balanced hazard ratio, $BHR = \exp(\beta)$ is simply the multiplicative factor to get the hazard of the high risk from the global hazard or from the low risk to the global one.

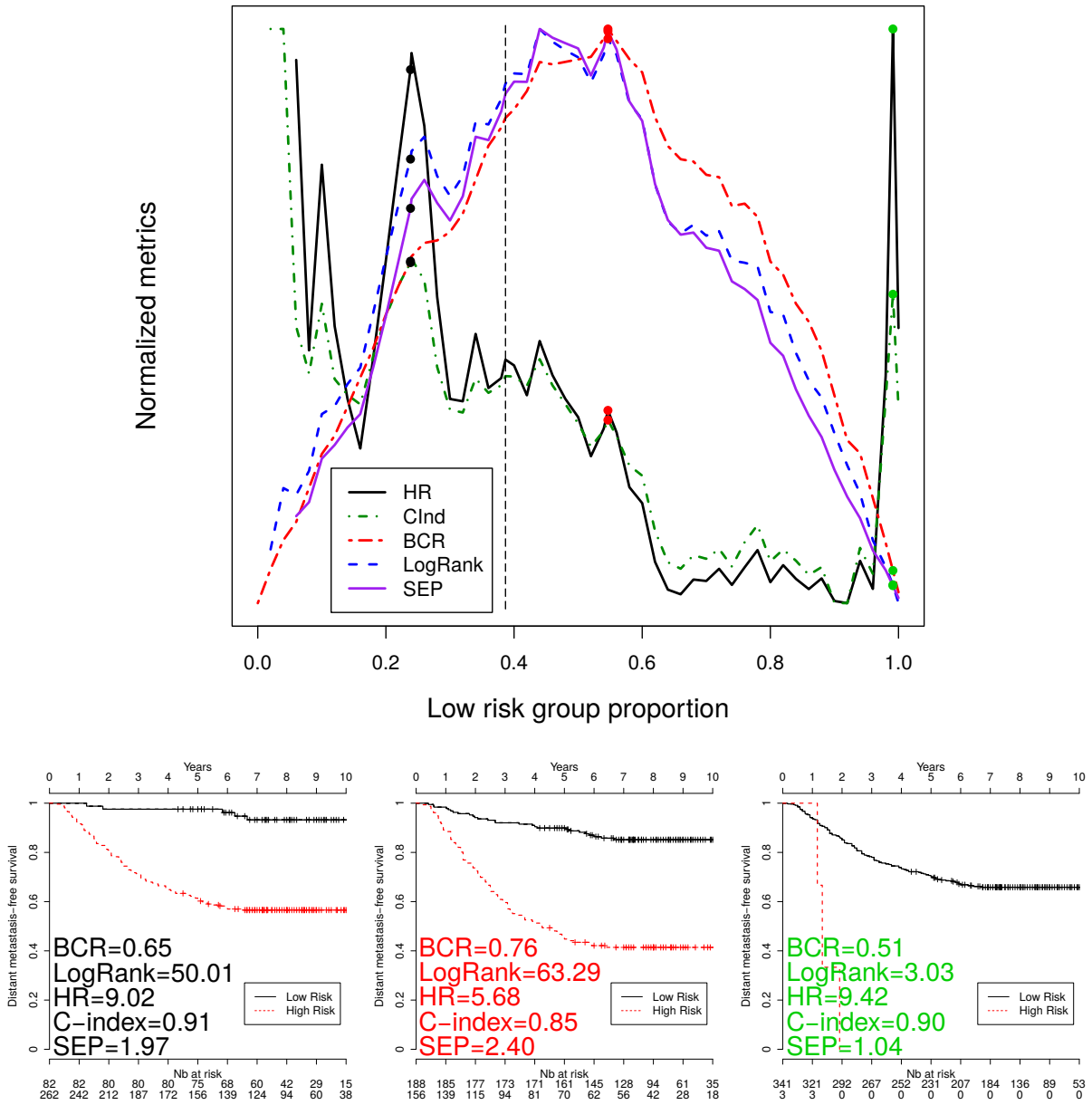
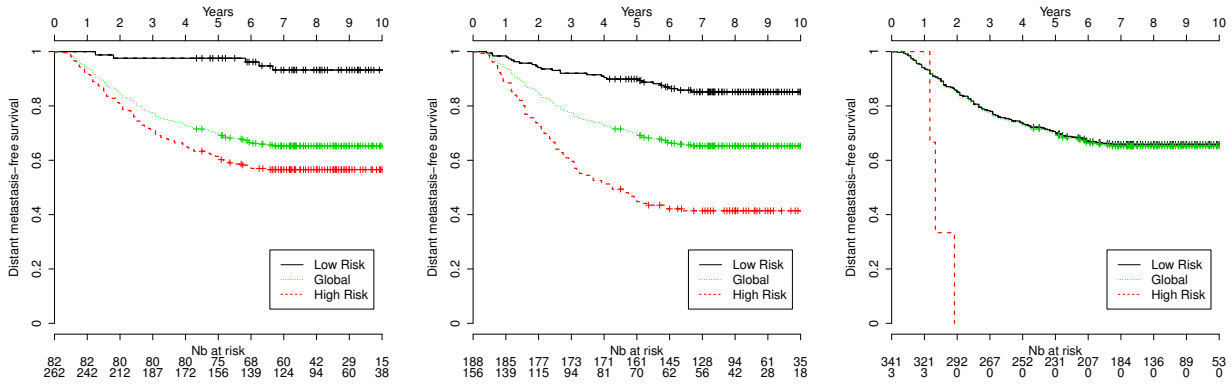


Figure 4. The various performance metrics evolution while varying the proportions in each risk group through adjusting the cut-off on risk scores. The original Gene76 cut-off corresponds to the dashed line.

Yet, according to the BHR formulation, whenever the vast majority of patients are artificially considered in one group, the difference between the global survival and the survival of this group will be small (see, for example, Fig. 5(a) and Fig. 5(c)). As such, the BHR penalized extremely unbalanced risk groups without forcing risk groups to be of equal size. The consideration of the global survival curve also has a smoothing effect on the BHR because a change in survival times for one specific group only affects the hazard ratio between this group and the global survival curve.

Figure 6 (a) further details our running example while reporting the BCR, logrank, HR and BHR (we left out the C-index and SEP for clarity as they behave like the HR and the logrank, respectively). The BHR exhibits a behavior similar to the BCR and the logrank while offering a natural interpretation in terms of survival differences between groups as the original HR. In particular, BHR is much smoother than HR, exhibits a data-dependent global optimum and penalizes artificially



(a) (b) (c)

Figure 5. Risk groups with 3 different cutoffs and the global survival curve (in green) for the Gene76 model on the VDX dataset.

unbalanced groups.

Unlike the logrank and the SEP, the BHR is sensitive to the risk group ordering: inverting the risk groups would lead to a BHR value below 1 while BHR tends to 1 when the survival differences between the risk groups tend to vanish.

The estimation $\hat{\beta}$ of the β value from the balanced hazard ratio (see equation (8)) is computed through the maximization of a partial likelihood, similarly to the original HR [14]. For the BHR, the partial likelihood is slightly modified to include the global survival without duplicating the patients. The partial likelihood for the balanced hazard ratio (with the standard Breslow approximation for ties [15]) is:

$$L = \prod_{i=1}^{n} \left[\frac{\exp(\beta g_i)}{\left(\sum_{j \in R(t_i)} \exp(\beta g_j) + 1 \right)^2} \right]^{\delta_i}, \text{with } g_i = -1 \text{ or } 1 \tag{9}$$

The $\hat{\beta}$ value, which is a maximum likelihood estimate, has an asymptomatically normal distribution. The variance of $\hat{\beta}$ is the inverse of the Fisher information [14] and can be estimated through the second derivative of the log-likelihood with respect to β:

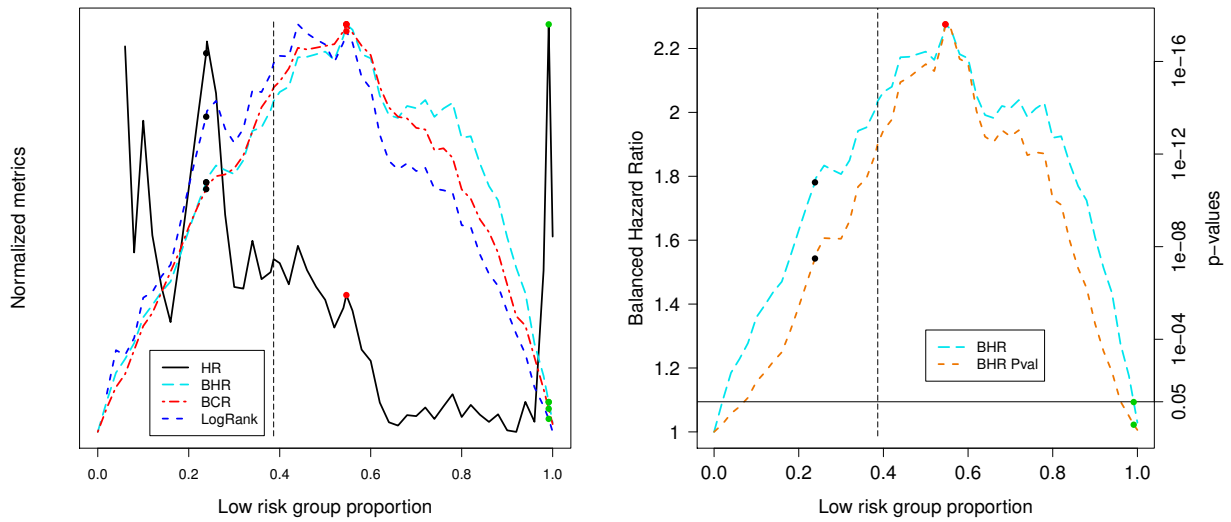
$$var(\hat{\beta}) \approx - \left(\frac{d^2 logL(\hat{\beta})}{d\beta^2} \right)^{-1} \tag{10}$$

One can thus easily compute confidence intervals and use standard statistical tests (Wald, likelihood ratio, score test [15]) to assess to which extent the BHR significantly departs from 1.

Figure 6 (b) illustrates that the associated p-values are fully concordant with the BHR values: an increase of BHR goes along a decrease of the associated p-value (here plotted in $-\log$ scale). Those results drastically contrasts with those obtained for the original HR (see figure 3).

7. BHR generalized to more than two risk groups

The original balanced hazard ratio (see equation (8)) is formulated with 2 risk groups and an additional global group representing all patients. The BHR can be easily generalized to an arbitrary number of original risk groups. The k original



(a) (b)

Figure 6. (a) Performance metrics on the VDX dataset while varying the proportions in each risk group. The original Gene76 cut-off corresponds to the dashed line. (b) Dashed orange line: p-values associated to BHR plotted in $-\log$ scale (right y-axis, the higher the better on such a plot)

risk groups are assumed to be ordered from lower to higher risk and arbitrarily numbered $1, 3, \ldots 2k-1$. For each pair (i, $i+2$) of consecutive risk groups following this order, one additional risk group numbered $i+1$ is considered gathering the patients of this pair of groups. In total, one considers $2k-1$ (original and additional) risk groups. The hazard function is now defined as follows:

$$h_i(t) = h_0(t)\exp(\beta g_i), \text{with } g_i \in [1:2k-1] \tag{11}$$

The balanced hazard ratio, $BHR = \exp(\beta)$ is the multiplicative factor between the hazards of two consecutive groups i and $i+1$. This formula is equivalent[‡] to the original BHR definition whenever $k=2$.

The SEP and BHR metrics are similar when restricted to 2 groups as they both compare the survival of each risk group to the global survival (yet only BHR is sensitive to the group ordering). When considering a larger and arbitrary number k of groups, those metrics differ more strongly since the BHR introduces multiple new groups while SEP only compares the survival in each (arbitrarily ordered) risk group to the global survival.

The evaluation of the 3 risk groups defined by the original NPI index (see section 3) is now straightforward according to this new BHR definition. The BHR can also be extended to continuous risk scores. Indeed, the limit case consists in assigning one single patient to each risk group. Each risk group would then be representative of a specific risk score and those groups can be sorted accordingly. While this is a natural limit case, it does not offer specific advantages over the HR computed on those risk scores. The purpose of introducing the BHR is to address the problems of the original HR whenever continuous risk scores are discretized into risk groups, as discussed in section 4.

8. Cut-off choice and risk group prediction

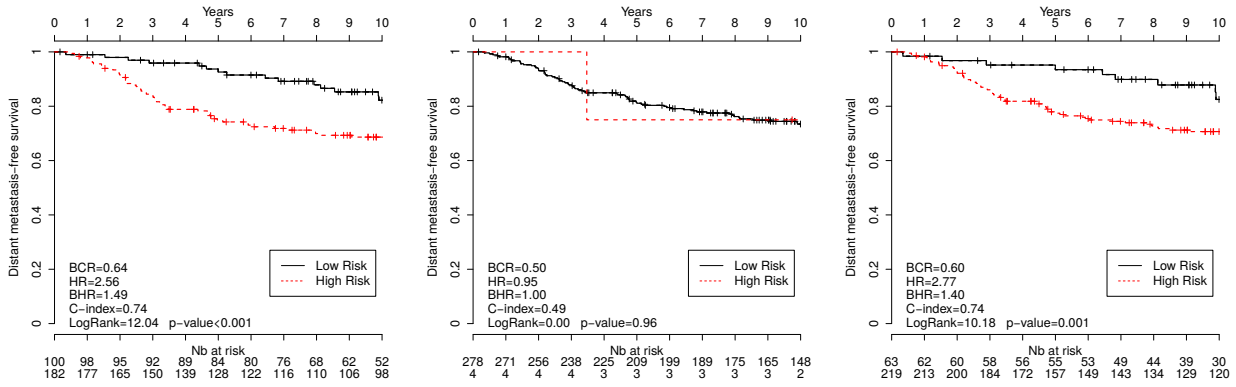
A relevant cut-off value is necessary to define risk groups from continuous risk scores, as formalized in equation (2). The choice of a specific cut-off is not always clearly motivated in the literature and sometimes not even explicitly described.

[‡] up to an arbitrary shift in group numbering.

However, given its potentially critical effect on the estimated quality of a prognostic model and its associated prognostic markers, it looks important to use an appropriate methodology to fix cut-offs. For instance, the original cut-off associated to the Gene76 prognostic index [3] has been chosen to achieve 100% of sensitivity and the highest specificity on the training set (an undisclosed fraction of the VDX dataset). Whether such choice is really optimal for the accurate prognosis on independent samples requires an additional validation as detailed below. In any case, even on the VDX dataset from which the Gene76 model has been estimated, the dashed line on figure 6 illustrate that such cut-off is arguably sub-optimal. Besides, choosing among all possible cut-offs the one minimizing the associated p-value underestimates such p-value due to the multiplicity of the test [8].

The BHR measure offers a natural way to choose a relevant cut-off and the associated risk groups. One typically considers a training and a validation set. Such validation set can be made of independent samples from the same study or, preferably, from an independent study on the same medical question. Such a scheme can even be generalized to a cross-validation protocol or while using several independent resamplings from various related clinical studies. In any case, the training data should typically be used to estimate the parameters of a prognostic model (*i.e.* the identity of the prognostic markers and the way to combine their values in a single risk score) as well as the cut-off on the risk scores. In particular, unless there is some prior knowledge on the relative size of each risk group, we recommend to choose the cut-off maximizing the BHR on the training set.

We illustrate the proposed methodology with the Gene76 model while changing its cut-off to maximize BHR on the VDX data from which it was originally estimated. We compare the predictive performances obtained on independent samples for various cut-off choices made on the training data. The BHR cut-off precisely corresponds on the training set (here the VDX data) to the risk group proportion defined by the red dot on figure 6. The original Gene76 cut-off is represented by the dashed line while the cut-off optimizing HR is represented by a green dot.



(a) (b) (c)

Figure 7. Prognostic performances on an independent validation set (TBG and UNT datasets) according to various cut-off choices on the training (VDX): largest BHR (a), largest HR (b), and the original Gene76 cut-off (c).

Figure 7 illustrates the impact of those 3 cut-off choices on independent samples for breast cancer prognosis. Their predictive performances were assessed in terms of BCR, hazard ratio, balanced hazard ratio, C-index and logrank. We consider in particular the TBG [9] and UNT [10] clinical studies as an independent validation set.

Figure 7(a) reports the survival curves of the risk groups resulting from Gene76 with the cut-off optimizing BHR on the training. Such choice leads to the best validation results. In contrast, choosing the cut-off to optimize the HR on the training leads to artificially unbalanced groups both on the training and on the validation set (Figure 7(b)) and much degraded validation performances. Figure 7(c) reports the validation results of the Gene76 model using its original cut-off. Those results are sub-optimal as well and illustrate that the Gene76 model could be made more effective by our proposed methodology to define risk groups. In particular the original cut-off leads to lower BCR, BHR and logrank values, which

are the 3 metrics not favoring artificially unbalanced risk groups.

The proposed methodology to fix a cut-off maximizing BHR on the training set is further validated with other prognosis models and other cancer studies in the supplementary materials. In all cases, those results illustrate the benefits on independent validation samples of considering BHR instead of HR for fixing those cut-off values. This methodology is also assessed on controlled experiments described in the next section.

9. Controlled experiments

The results presented in section 8 show that maximizing the BHR to choose a cut-off on risk scores is a good approach to optimize prognosis performances on new and independent samples. We further motivate this approach in controlled experiments for which an underlying threshold between risk groups is *a priori* fixed according to a prescribed proportion between a high or low risk profile. We assess in particular to which extent the proposed methodology is able to recover the appropriate proportions between risk groups and hence the underlying cut-off to be set on risk scores.

Synthetic data are generated for n patients separated into a low $(g_i = -1)$ or high $(g_i = 1)$ risk group according to prescribed proportions, ρ and $1-\rho$ respectively. Survival data and risk scores (t_i, δ_i, r_i) are randomly generated for each sample $i \in [1, n]$. The time to event t_{e_i} of patient i is drawn from a Weibull distribution and the censoring time t_{c_i} is drawn from an exponential distribution:

$$t_{e_i} \sim \text{Weibull}(\lambda_1 \exp(-\frac{\mu g_i}{2k}), k) \tag{12}$$

$$t_{c_i} \sim \text{Exp}(\lambda_2) \tag{13}$$

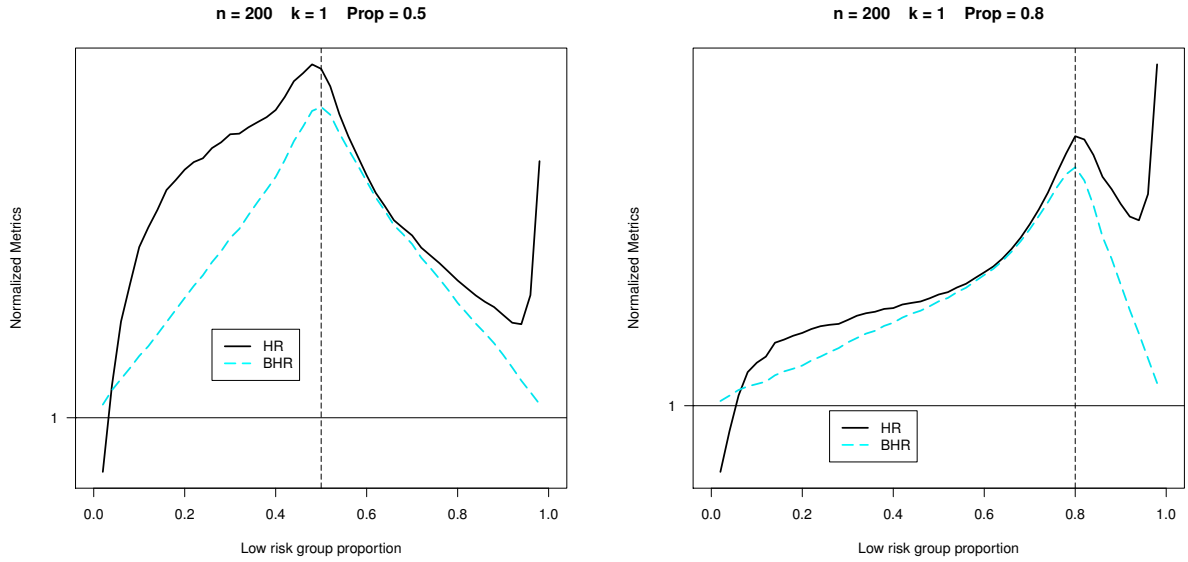
The scaling parameters $\lambda_1 = 0.002$ and $\lambda_2 = 0.01$ are fixed according to [16] from which these experiments are inspired. The shape parameter k varies in $[0.5, 1.5]$ in our experiments. Patient i is censored $(\delta_i = 0)$ if his censoring time t_{c_i} occurs before the time to event t_{e_i} and the time t_i is simply defined as the minimum between both times: $t_i = \min(t_{e_i}, t_{c_i})$. The true hazard ratio between groups can be directly controlled using this protocol since it is given by $HR = \exp(\mu)$.

The risk score of a patient i is drawn from a Normal distribution centered on g_i, -1 or 1, respectively for low or high risk group. The risk scores are then distributed according to a mixture of the two Normal distributions, according to the prescribed proportion ρ between risk groups: $r_i \sim \rho\mathcal{N}(-1, 0.5) + (1-\rho)\mathcal{N}(1, 0.5)$. The underlying threshold to be discovered is defined as the ρ-percentile of this distribution. We note that a perfect discrimination between risk groups could hardly be obtained since risk scores overlap across risk groups, as expected in a real scenario. Results are reported below over 500 independent runs of such controlled experiments.

Figure 8 reports the hazard ratio (HR) and balanced hazard ratio (BHR) computed while varying the cut-off used to define the risk groups. Results are reported here for two prescribed proportions $\rho = 50\%$ or $\rho = 80\%$ of the low risk profile but the same conclusions can be drawn from other ρ values. In particular, maximizing BHR leads to chose a cut-off on risk scores which, when averaged over 500 runs, corresponds to the correct underlying proportion between risk groups. In contrast, maximizing HR may lead to an inappropriate cut-off choice favoring strongly unbalanced groups.

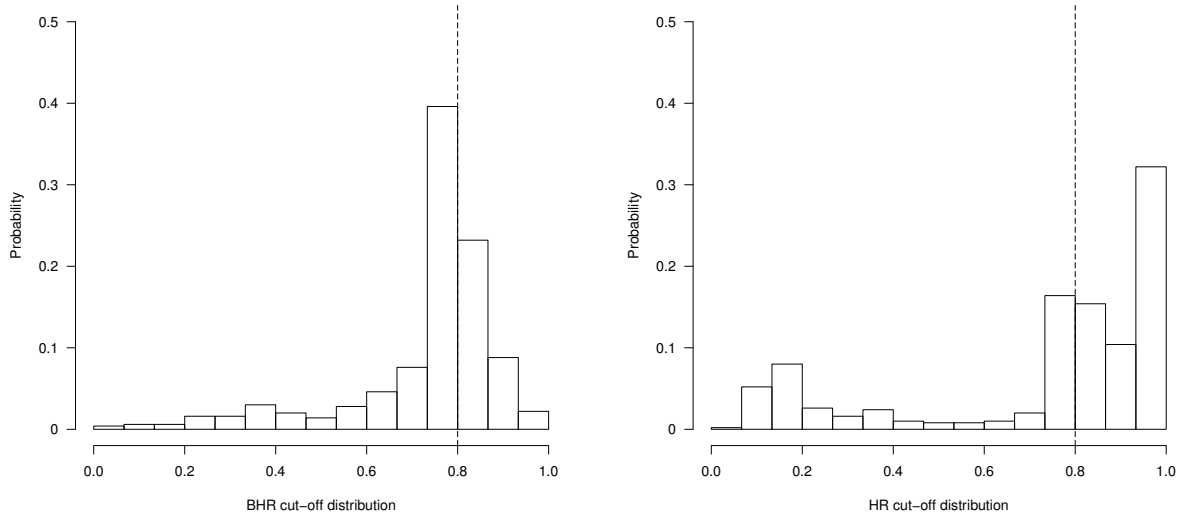
Figure 9 offers a closer look at the distribution over these 500 runs of the proportions between risk groups for which BHR, respectively HR, is maximum. The true proportion ρ in the low risk group was here fixed to 80%. The maximal BHR is clearly more concentrated around the true underlying proportion while the maximum HR distribution is much more dispersed and skewed towards an excessively large value.

Figure 10 generalizes the above analysis while changing the shape parameter k of the Weibull distribution used to generate the survival data. It illustrates that selecting a cut-off value while maximizing HR would be even more inappropriate as k is increased to 1.5 (see, in particular, Figure 10 (b)). In contrast, maximizing BHR remains an



(a) (b)

Figure 8. Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, the shape parameter $k = 1$ and the true group hazard ratio $\exp(\mu) = 3$. The true proportion ρ of patients in the low risk group was set to 50% (Figure 8 (a)) or 80% (Figure 8 (b)).

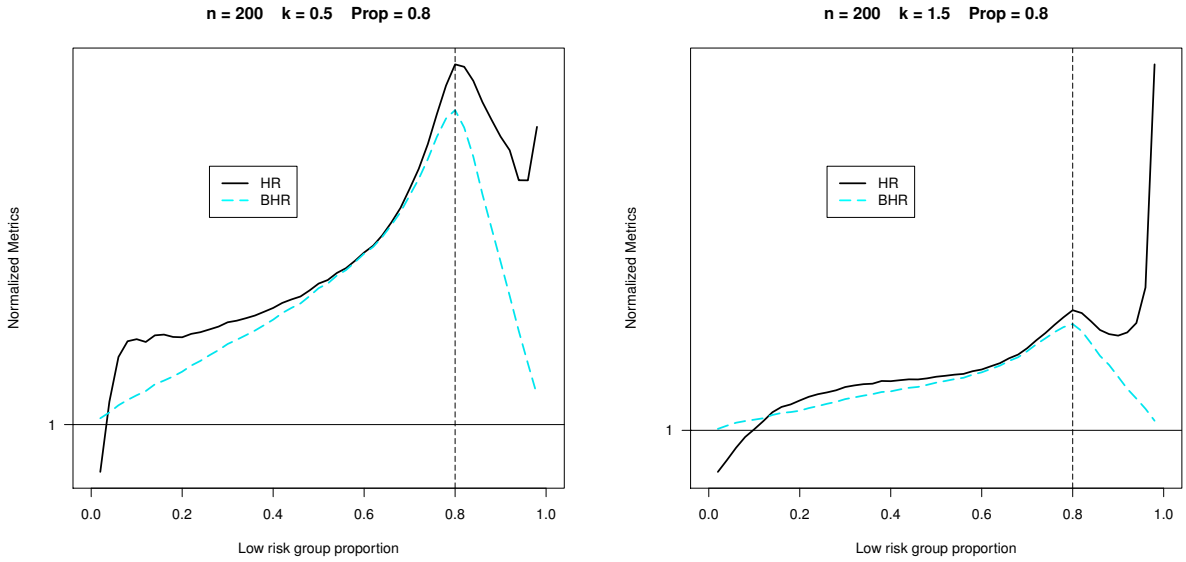


(a) (b)

Figure 9. Distribution over 500 runs of the low risk group proportion for which BHR (a)) respectively HR (b), is maximum. The experiments are conducted with $n = 200$ patients, the shape parameter $k = 1$ and the true group hazard ratio $\exp(\mu) = 3$. The true underlying proportion of patients in the low risk group was set to 80%.

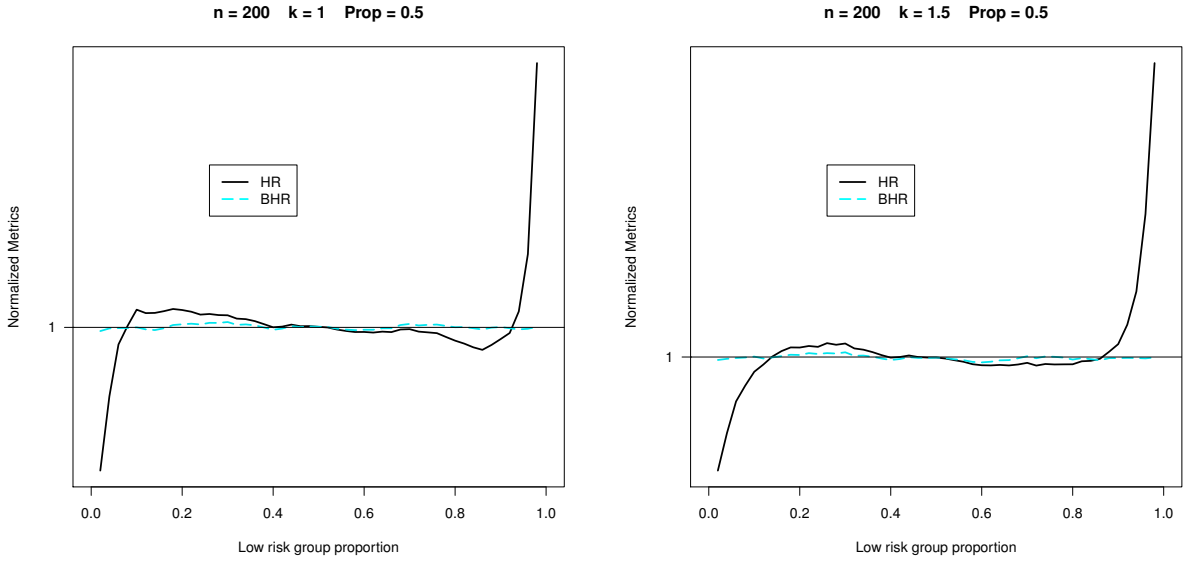
appropriate methodology across various shape values.

Our final experiments is considering a true group hazard ratio $\exp(\mu)$ equal to 1. In other words, the data is generated such that there is actually no survival differences between both groups. Figure 11 reports the averaged BHR and HR values over 500 runs while changing the cut-off on risk scores. The flat BHR curve illustrates that no specific cut-off should be chosen here and hence all patients should be assigned to a common risk group. In contrast, when maximizing HR, a dichotomization into highly unbalanced groups is again promoted.



(a) (b)

Figure 10. Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, the true proportion $\rho = 0.8$ in the low risk group and the true group hazard ratio $\exp(\mu) = 3$. The shape parameter k of the Weibull distribution was set to 0.5 (a) or 1.5 (b).



(a) (b)

Figure 11. Evolution of HR and BHR, averaged over 500 runs, while varying the proportions in each risk group through adjusting the cut-off on risk scores. The experiments are conducted with $n = 200$ patients, $\rho = 0.5$ and a true group hazard ratio $\exp(\mu)$ equal to 1 (no survival difference between groups). The shape parameter k of the Weibull distribution was set to 1 (a) or 1.5 (b).

10. Conclusion and perspectives

Properly assessing risk group prediction from survival data is essential to analyze the relevance of candidate prognosis markers. We show here that the group hazard ratio (HR) and the concordance index, often used in such a context, can be inappropriately optimized by considering artificially unbalanced risk groups. They also exhibit many local optima and

non smooth behaviors which make their evaluation highly sensitive to small fluctuations. Alternative existing metrics include the logrank statistics, the SEP metric and the average between sensitivity and specificity, also called BCR. While the logrank is harder to interpret in terms of survival differences between risk groups, the BCR is not fully adequate to censored data and relies on an additional critical timepoint that is highly dependent on the pathology. Moreover, the logrank and the SEP are insensitive to a specific ordering of the risk groups. They measure survival differences between risk groups but not the orientation of those differences.

We introduce here the balanced hazard ratio (BHR) which has a similar interpretation as the original HR, arguably the most common metric used by clinicians and bio-statisticians to assess risk groups. The BHR penalizes extremely unbalanced risk groups and, more generally, offers a smoother profile with a natural optimum. Its value is much less influenced by marginal changes in the proportions between risk groups and its behaves consistently with its associated p-value.

We further show how the BHR may be generalized to an arbitrary number of risk groups and how it can be used to choose an appropriate cut-off on risk scores to define risk groups. Such a methodology is both simple computationally and is shown to be sound in controlled experiments as well as real clinical studies.

Our future work includes the design of estimation algorithms for prognosis models optimizing BHR directly. Currently, Cox hazards models are typically estimated to optimize the partial likelihood on the training data. It looks preferable to fit parameters optimizing directly the final performance metric of the prognostic model. We have shown here that optimizing the group hazard ratio would be largely inappropriate. Fitting a model to optimize the BHR instead looks to be a promising alternative.

The balanced hazard ratio could also be extended to competing risks. Competing risks are modeled either through cause-specific hazards [17] or cumulative incidence functions [18]. The other causes of events are dealt with in such models either through censoring or as events occurring at an infinite time. Similarly to our proposal, a third global risk group could be introduced in these models to penalize extremely unbalanced risk groups.

Acknowledgement

Funding The work of Samuel Branders was supported by the F.R.S. - FNRS - Télévie (Grant number FC 88088) and by the Fondation Louvain.

References

1. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast cancer research and treatment* Jan 1992; **22**(3):207–19. URL http://www.ncbi.nlm.nih.gov/pubmed/1391987.
2. Hoster E, Dreyling M, Klapper W, Gisselbrecht C, van Hoof A, Kluin-Nelemans HC, Pfreundschuh M, Reiser M, Metzner B, Einsele H, *et al.*. A new prognostic index (MIPI) for patients with advanced-stage mantle cell lymphoma. *Blood* Jan 2008; **111**(2):558–65, doi:10.1182/blood-2007-06-095331. URL http://www.ncbi.nlm.nih.gov/pubmed/17962512.
3. Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, *et al.*. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005; **365**(9460):671–9, doi:10.1016/S0140-6736(05)17947-1. URL http://www.ncbi.nlm.nih.gov/pubmed/15721472.
4. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, *et al.*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine* Dec 2004; **351**(27):2817–26, doi:10.1056/NEJMoa041588. URL http://www.ncbi.nlm.nih.gov/pubmed/15591335.
5. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AaM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al.*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* Jan 2002; **415**(6871):530–6, doi:10.1038/415530a. URL http://www.ncbi.nlm.nih.gov/pubmed/11823860.

6. Cox D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B* 1972; **34**(2):187–220. URL http://www.jstor.org/stable/10.2307/2985181.
7. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* 1996; **15**:361–387.
8. Altman DG. Categorising continuous variables. *British journal of cancer* 1991; **64**(5):975.
9. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, Viale G, Delorenzi M, Zhang Y, D'Assignies MS, *et al.*. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research : an official journal of the American Association for Cancer Research* Jun 2007; **13**(11):3207–14, doi:10.1158/1078-0432.CCR-06-2765. URL http://www.ncbi.nlm.nih.gov/pubmed/17545524.
10. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, *et al.*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* Feb 2006; **98**(4):262–72, doi:10.1093/jnci/djj052. URL http://www.ncbi.nlm.nih.gov/pubmed/16478745.
11. Haibe-Kains B, Desmedt C, Rothé F, Piccart M, Sotiriou C, Bontempi G. A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome biology* Jan 2010; **11**(2):R18, doi:10.1186/gb-2010-11-2-r18. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872878&tool=pmcentrez&rendertype=abstract.
12. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.
13. Sauerbrei W, Hübner K, Schmoor C, Schumacher M. Validation of existing and development of new prognostic classification schemes in node negative breast cancer. *Breast cancer research and treatment* 1997; **42**(2):149–163. URL http://link.springer.com/article/10.1023/A:1005733404976.
14. Cox DR, Hinkley DV. *Theoretical statistics*. CRC Press, 1979.
15. Collett D. *Modelling survival data in medical research*. Chapman&Hall-CRC, 2003.
16. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in medicine* Mar 2004; **23**(5):723–48, doi:10.1002/sim.1621. URL http://www.ncbi.nlm.nih.gov/pubmed/14981672.
17. Prentice R, Kalbfleisch J. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; **34**(4):541–554. URL http://www.jstor.org/stable/2530374.
18. Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical* ... 1999; **94**(446):496–509. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144.