# Dichotomization of survival variables in medical prediction problems

**Marion Verduijn**[1,3], **Niels Peek**[1] and **Evert de Jonge**[2]

[1] Dept. of Medical Informatics, [2] Dept. of Intensive Care Medicine

Academic Medical Center, University of Amsterdam,

P.O. Box 22700, 1100 DE Amsterdam, The Netherlands

[3] Dept. of Biomedical Engineering, University of Technology, Eindhoven, The Netherlands

e-mail: {m.verduijn, n.b.peek, e.dejonge}@amc.uva.nl

## Abstract

Problems with modeling survival outcome arise when this outcome is censored informatively. Dichotomization of the survival outcome is a possible solution for this problem. However, no consensus exists on the threshold selection procedure. This paper argues that the threshold should be selected in a structured fashion. It presents a dichotomization method that builds on the notion of separability of outcome classes, which is quantified in term of the Bayes error rate. The approach is applied on the problem of predicting length-of-stay at the intensive care unit.

## 1 Introduction

A possible approach to building a prognostic model for *survival outcomes*, such as time until death and length of hospitalization, is by conducting survival analysis [9] in the form of developing a Cox proportional hazards model [7]. However, when the outcome is censored by an event that is strongly related to the survival time itself, which is, for instance, the case when length of hospitalization is censored by death, the standard Cox model is not suitable: this *informative censoring* can lead to parameter estimates that are seriously biased and false predictions [1].

A solution for modeling this type of survival outcome is dichotomization of the variable; this reduces the prediction problem to a classification problem with two categories allocating all censored cases to the worst outcome class. The loss of information in the outcome variable may be compensated for by increased possibilities to build a reliable prognostic model.

The dichotomization of survival variables is frequently applied in medical prediction problems [12; 13; 14]. However, the method of dichotomization is often *ad hoc* (e.g. based on the 95% percentile) when no clinical knowledge is available to do so. This is unfortunate as suboptimal threshold selection can lead to low accuracy of the prognostic model that is developed for the dichotomized outcome and to restricted insight into the structure of the prediction problem. In this paper, we argue that threshold selection should be conducted in a structured fashion, and be based on the available data as much as possible. We present two structured methods to select the threshold for dichotomization, based on *separability* [10] of the resulting outcome classes which is quantified in terms of the Bayes error rate [2].

The paper is organized as follows. Section 2 describes methods to dichotomize based on separability measurement. These methods are applied to the problem of predicting length of stay in the intensive care unit (ICU) in Section 3. We discuss the results in Section 4.

## 2 Separability measurement

The degree of difficulty of a classification problem depends on how well the outcome classes can be separated from each other. The concept of class separability stems from *Bayesian decision theory* [2]. First, we will introduce some notation. Let the two classes of the dichotomized outcome variable be denoted by $\omega_1$ and $\omega_2$ and let $P(\omega_j)$ denote the prior probability of class $\omega_j$, $j = 1, 2$. Furthermore, let $p(\mathbf{x})$ denote the probability density function, and let a class-conditional probability be denoted by $P(\omega_j|\mathbf{x})$, where $\mathbf{x}$ is a covariate vector with $k$ different features $x_1, \cdots, x_k$.

### 2.1 Bayes error rate

In Bayesian decision theory, the task is to design a decision rule that assigns a class to a given object $\mathbf{x}$. For minimizing the probability of error, the *Bayes decision rule* says: Decide $\omega_1$ if $P(\omega_1 \,|\, \mathbf{x}) > P(\omega_2 \,|\, \mathbf{x})$; otherwise decide $\omega_2$. Under this rule, the probability of misclassifying $\mathbf{x}$ becomes

$$P_{Bayes}(error|\,\mathbf{x}) = \min\{P(\omega_1 \,|\, \mathbf{x}), P(\omega_2 \,|\, \mathbf{x})\}. \quad (1)$$

The *Bayes error rate* (BER) is defined as

$$P^* = \int P_{Bayes}(error|\,\mathbf{x})p(\mathbf{x})\, d\mathbf{x}, \quad (2)$$

and provides a lower bound on the error rate that may be achieved by any decision rule [8]. If $P^* = 0$, then the two classes are completely separable; higher values represent increasing inseparability of the classes, and indicate that we are dealing with increasingly difficult problems. Therefore, assessing the BER for outcomes dichotomized using different thresholds can be used to select the threshold that best separates the outcome classes.

In practice, the assessment of the BER is hindered because not all probability values that are relevant to

the problem are known, but need to be estimated from a finite sample. Furthermore, even if these probabilities are known, full calculation of the BER requires the computing of a $k$-fold, multiple integral. Below, we describe two approaches to solve these problems by approximating the BER.

### 2.2 Approximation of upper bounds

The first approach to overcome difficulties related to BER assessment is based on deriving upper bounds of the BER. We discuss three theoretical upper bounds of the right-hand side of Equation 2 and describe how they are computed in practice.

The first upper bound is derived by simply neglecting all information from the covariate vector $\mathbf{x}$ and focusing on the prior probabilities of both outcome classes. This results in the decision rule 'Decide $\omega_1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide $\omega_2$', which has an associated error rate of

$$P^*_{prior} \quad = \quad \min\{P(\omega_1), P(\omega_2)\}. \tag{3}$$

It is easily verified that $P^* \leq P^*_{prior}$. Furthermore, it can be proved that $P^* = P^*_{prior}$ in case of highly unbalanced prediction problems (i.e. one of the prior probabilities $P(\omega_1)$ and $P(\omega_2)$ is very small). In practice, the prior probabilities can be estimated reliably from a relatively small number of observations. The bound $P^*_{prior}$ is therefore simple and reliable, but probably not very effective when much information resides in the covariate vector $\mathbf{x}$.

More sophisticated upper bounds are obtained by the methods of H. Chernoff [5] and A. Bhattacharyya [3], for which the assumption is made that the observations are drawn from class-conditional normal densities. The *Chernoff bound* has the form

$$P^*_{Cher.} = P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)e^{-k(\beta)} \tag{4}$$

for $0 \leq \beta \leq 1$ where

$$k(\beta) = \frac{\beta(1-\beta)}{2}\ \alpha\ +\ \frac{1}{2}\ln\frac{|\beta\mathbf{\Sigma}_1 + (1-\beta)\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|^{\beta}|\mathbf{\Sigma}_2|^{1-\beta}}$$

and

$$\alpha = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t[\beta\mathbf{\Sigma}_1 + (1-\beta)\mathbf{\Sigma}_2]^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

In this equation, $\boldsymbol{\mu}_j$ and $\mathbf{\Sigma}_j$ are respectively the mean vector and the covariance matrix of $\mathbf{x}$ for class $\omega_j$. To compute the *Chernoff bound* on $P^*$, we have to choose the parameter $\beta$ such that $e^{-k(\beta)}$ is minimal, and substitute the results in Equation 4. The proper value for $\beta$ is easily found by numerical optimization.

A computationally simpler but slightly less tight bound can be derived by simply setting $\beta = 0.5$. This is called the *Bhattacharyya bound*. This bound can be used instead of the Chernoff bound when the majority of class separation comes from differences in class means and (co)variances because the prior class probabilities are roughly equal, i.e. $P(\omega_1) \simeq P(\omega_2)$.

For calculation of the Chernoff and Bhattacharyya bounds in practice, not only the prior probabilities $P(\omega_1)$ and $P(\omega_2)$, but also the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ have to be estimated from data and plugged into the relevant formulas. Note that a $d \times d$ covariance matrix has $\frac{1}{2}d \cdot (d-1)$ independent entries, and that each of these entries is estimated from a subset of the data that corresponds to one outcome class. These estimates may quickly become unreliable with a growing number of features $d$, or when the problem is highly unbalanced. Furthermore, when the assumption of normality of the class-conditional densities is violated, the computed bounds may be false.

### 2.3 Building predictive models

Determining the BER requires knowledge of the class-conditional probability values $P(\omega_j \mid \mathbf{x})$, which are in practice unknown. A second approach to estimating these probability values consist of developing a predictive model from the data at hand. The potential disadvantage of this approach is that the results may be highly dependent on the modeling method. If the chosen method is unable to separate the two outcome classes, this may be related to the difficulty of the classification problem, but also to limitations of the method.

Here, we have used two methods to develop predictive models: *kernel smoothing* [6] and *regression trees* [4]. Kernel smoothing is a nonparametric regression method that avoids lack-of-fit problems by making very few assumptions on the characteristics of the classification problem. A disadvantage of kernel smoothing that is provides very little insight into the structure of problem; however, this is not a problem here as we are only interested in finding the optimal dichotomization threshold. The regression tree method belongs to the family of recursive partitioning techniques and was chosen here because it represents a popular choice of modeling method in the literature on classification problems. Both methods are briefly described below.

For the BER estimation using predictive models, the probability $P(\omega_j \mid \mathbf{x})$ is estimated for every observation $\mathbf{x}$ in the data set. The computation of the $k$-fold, multiple integral is replaced by a summation over each of these observations. This approach approximates the BER itself. Theoretically, these approximations are also upper bounds, because no classification could improve over the Bayes decision rule. However, due to instable estimations and overfitting, the BER can also be underestimated.

## 3 Application to ICU data

We have applied the methods for dichotomization to the outcome *length of stay at intensive care unit* (ICU LOS) after cardiac surgery. This outcome can be used as a proxy for the 'degree of complication' during the recovery process and therefore, as a measure of the quality of delivered care. Prediction of this outcome helps doctors to get insight into the complication risk of their patients and supports ICU managers in planning the ICU capacity. ICU LOS is informatively censored by death, because this most serious complication that may occur, breaks off the stay at the ICU. Therefore, we have determined a threshold for dichotomization based on a data set, allocating all patients who died to patients with an ICU LOS longer than the threshold value. So, two outcome categories have been created: *short LOS*, and *long LOS or death*.

Table 1: Approximations of the Bayes error rate $P^*$ for dichotomized outcomes of ICU LOS.

| threshold | # events | prior bound | Bhattacharyya bound | Chernoff bound | kernel smoothing | | regression tree | |
|---|---|---|---|---|---|---|---|---|
| | | $P^*_{prior}$ | $P^*_{Bhat.}$ | $P^*_{Cher.}$ | $P^*_{kernel}$ | AUC | $P^*_{tree}$ | AUC |
| 2 days | 1137 | 0.449 | 0.464 | 0.459 | 0.385 | 0.681 | 0.324 | 0.684 |
| 3 days | 818 | 0.397 | 0.448 | 0.456 | 0.339 | 0.703 | 0.260 | 0.704 |
| 4 days | 623 | 0.302 | 0.417 | 0.433 | 0.268 | 0.712 | 0.232 | 0.730 |
| 5 days | 530 | 0.257 | 0.395 | 0.415 | 0.225 | 0.719 | 0.197 | 0.733 |
| 6 days | 443 | 0.215 | 0.371 | 0.394 | 0.191 | 0.726 | 0.171 | 0.741 |
| 7 days | 385 | 0.187 | 0.348 | 0.374 | 0.167 | 0.733 | 0.153 | 0.742 |
| 8 days | 342 | 0.166 | 0.327 | 0.350 | 0.147 | 0.757 | 0.137 | 0.749 |
| 9 days | 320 | 0.155 | 0.319 | 0.341 | 0.140 | 0.757 | 0.128 | 0.754 |
| 10 days | 290 | 0.141 | 0.303 | 0.326 | 0.126 | 0.767 | 0.118 | 0.703 |
| 12 days | 254 | 0.123 | 0.286 | 0.311 | 0.108 | 0.774 | 0.108 | 0.692 |
| 14 days | 224 | 0.109 | 0.272 | 0.298 | 0.091 | 0.776 | 0.094 | 0.672 |
| 16 days | 206 | 0.100 | 0.259 | 0.284 | 0.084 | 0.785 | 0.092 | 0.689 |
| 18 days | 189 | 0.092 | 0.248 | 0.274 | 0.075 | 0.788 | 0.086 | 0.625 |
| 20 days | 177 | 0.086 | 0.239 | 0.263 | 0.070 | 0.789 | 0.078 | 0.681 |
| 25 days | 159 | 0.077 | 0.225 | 0.241 | 0.062 | 0.789 | 0.074 | 0.617 |
| 30 days | 140 | 0.068 | 0.209 | 0.222 | 0.052 | 0.802 | 0.063 | 0.754 |
| 40 days | 121 | 0.059 | 0.190 | 0.203 | 0.045 | 0.816 | 0.057 | 0.683 |
| 50 days | 116 | 0.056 | 0.188 | 0.201 | 0.043 | 0.811 | 0.052 | 0.699 |
| 70 days | 108 | 0.052 | 0.180 | 0.195 | 0.039 | 0.825 | 0.051 | 0.694 |
| 110 days | 103 | 0.050 | 0.173 | 0.186 | 0.038 | 0.836 | 0.047 | 0.660 |

# events: number of patients with ICU LOS higher than threshold value or death

### 3.1 Data and methods

A data set was used from cardiac operations conducted at the Academic Medical Center in Amsterdam, the Netherlands, in the years 1997–2001. Based on this data set, regression trees for the outcome mortality have been developed [15]. The data set contains 144 data items including patient characteristics such as age, surgery type and indicators of the patient's state during the first 24 hours at the ICU such as blood and urine values for 3855 patients. Because of including these latter data items, we have excluded all patients who left the ICU within one day. Based on data of the remaining 2063 patients, of which 101 patients died (4,9%), we have determined the separability of the dichotomized outcomes of ICU LOS.

We have quantified the separability of outcomes that are defined according to thresholds of 2 days up to and including 110 days by approximating the BER. Separability of outcomes with higher thresholds was not determined, because of the low proportion ($< 5\%$) of patients who stay more than 110 days or die. Note that from a threshold of 18 days onwards, the majority of patients in the category *long LOS or death* died, so that emphasis in the classification problem shifts towards predicting death instead of prolonged ICU stay.

The computation of the Chernoff and Bhattacharyya bounds and application of kernel smoothing require that a small number of continuous features be selected. We have selected the features *perfusion time* (the duration of using the heart-lung machine), *maximal creatinine value* and *minimal systolic blood pressure* based on univariate statistical analyses. The latter two variables are measured during the first 24 hours of the ICU stay. Because for calculation of the Bhattacharyya and Chernoff bounds normal class-conditional densities are assumed, we have used the logarithmic transformation of the features *perfusion time* and *maximal creatinine value*. The model building procedure of regression trees includes its own feature selection mechanism. We have evaluated the ability of the predictive models to discriminate between the two outcome classes using the area under the ROC curve (AUC). The approximations of the BER based on predictive models and the model evaluation measures were calculated using 10-fold cross validation.

### 3.2 Results

The results are summarized in Table 1. Each table row contains the threshold that is used for dichotomization, and the number of cases with an ICU LOS higher than the threshold value or death. Furthermore, approximations of the BER are described based on prior probabilities, the Bhattacharyya and Chernoff bounds, and using kernel smoothing and regression trees. For these predictive models, the AUC values are also described in the table.

It appears from this table that BER approximation using the Bhattacharyya and Chernoff bounds result in high values, compared to approximations made based on prior probabilities and using predictive models. Compared to the Chernoff bound, the lower values of the Bhattacharyya bound are unexpected, as this bound is theoretically less tight.

Kernel smoothing is found to be well able to separate the outcome classes when the number of events are low (threshold of 110 days: $P^*_{kernel}$ is 0.038, which is 24% error improvement compared to $P^*_{prior}$), while the regression tree only reduces 6% of the error ($P^*_{tree}$ is 0.047). This confirms the flexibility of kernel smoothing by fitting the data well, also when the number of events is low. Compared to kernel smoothing, we found that regression trees reduce more error in the case of low threshold values (3–10 days). This indicates that for more balanced classification problems regression trees provide better predictions probably due to the feature selection method.

We have found that beginning at the threshold of 10

days, the sensitivity of kernel smoothing is equal to 0. So, in these cases, using the decision rule 'Decide $\omega_1$ if $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$; otherwise decide $\omega_2$', kernel smoothing classifies all patients who stayed long at the ICU or who died into the outcome category *short LOS*. Combining these findings to the AUC values, kernel smoothing is found to discriminate well between both outcome classes, but only when using probability thresholds for classification that are at least lower than 0.5. Comparing this to the regression trees, we found lower AUC values, while the sensitivity is higher than 0 for all thresholds. So, for these cases, the regression trees give only the appearance of separation.

We can draw some conclusions from the approximated BER about the selection of threshold values for ICU LOS. When increasing the threshold value, the BER decreases to 0. For that reason, we have focused on the decrease of BER per increase of the threshold value instead of the approximation of the BER itself. We have found that the decrease of BER based on prior probabilities, the Bhattacharyya and Chernoff bounds, and using kernel smoothing is maximal between 3 and 4 days (a decrease of 0.095, 0.031, 0.023 and 0.071 respectively), while for the regression tree method the decrease is maximal between 2 and 3 days (a decrease of 0.064). In practice, the optimal value would possibly be somewhere between these values; insight into this value would be derived when using smaller scales to dichotomize (e.g. hours instead of days).

## 4 Discussion and conclusions

If threshold selection to dichotomize survival variables cannot be based on clinical knowledge, it should be based on data analysis conducted in a structural fashion. This paper describes approaches to select thresholds based on the notion of separability of outcome classes, which is quantified in terms of approximations of the BER. In the literature, additional measures of separability have been described [10; 11], such as *information value* (or *entropy*) and the related *Gini index*. These measures are not directly related to the BER. For a more complete investigation of methods to quantify separability of outcome classes, these methods should also be taken into account.

In the application described, the behaviour of the Bhattacharyya and Chernoff bounds is disappointing. This is probably due to the fact that when increasing the threshold, the estimations for outcome *long LOS or death* were based on a low number of cases by which the estimated covariance matrices are probably instable. This problem may partly be solved by making parametric assumptions about the covariance matrices.

The study is limited by the fact that we did not perform extensive analyses for feature selection, taking a possible shift of features related to higher threshold values into account; except for regression trees, one set of selected features was used for calculation of the measures for all dichotomized outcomes. However, each dichotomization threshold induces a new classification problem that possibly has its own optimal set of features.

In the future, we intend to continue the investigation using additional methods (e.g., entropy) and more types of methods to build predictive models. To confirm the described findings, we will perform method validation based an independent test set. Furthermore, we will investigate the influence of feature selection on the separability of the outcome classes.

**Acknowledgement**
The authors would like to thank Koos Zwinderman for his methodological suggestions.

## References

[1] P.D. Allison. *Survival analysis using the SAS System; a practical guide*. SAS Institute Inc., Cary, NC, USA, 1995.

[2] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[4] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, 1984.

[5] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[6] W.S. Cleveland and C. Loader. Smoothing by local regression: Principles and methods. *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49, 1996.

[7] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[8] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.

[9] T.R. Fleming and D.P. Harrington. *Counting processes and survival analysis*. John Wiley & Sons, New York, 1991.

[10] D.J. Hand. *Construction and assessment of classification rules*. John Wiley & Sons, New York, 1997.

[11] W.E. Pierson Jr. *Using boundary methods for estimating class separability*. PhD thesis, Ohio State University, 1998.

[12] J.P. Marcin, A.D. Slonim, M.M. Pollack, and U.E. Ruttimann. Long-stay patients in the pediatric intensive care unit. *Critical Care Medicine*, 29:652–657, 2001.

[13] P.K. Stein, R.E. Schmieg, A. El-Fouly, P.P. Domitrovich, and T.G. Buchman. Association between heart rate variability recorded on postoperative day 1 and length of stay in abdominal aortic surgery patients. *Critical Care Medicine*, 29:1738–1743, 2001.

[14] J.V. Tu, S.B. Jaglal, C.D. Naylor, and the Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery. *Circulation*, 91:677–684, 1995.

[15] M. Verduijn, N. Peek, and J.J. Kloek. Predicting outcomes of cardiac surgery: modeling the interaction of risk factors. In *Intelligent Data Analysis in Medicine and Pharmacology IDAMAP*, 2002.