Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes

Brent A. Williams, MS, Jayawant N. Mandrekar, PhD,
Sumithra J. Mandrekar, PhD,
Stephen S. Cha, MS, Alfred F. Furth, MS

Technical Report Series #79
June 2006

Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota

Copyright 2006 Mayo Foundation

TECHNICAL REPORT

Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes

Brent Williams, M.S.

Jayawant N. Mandrekar, Ph.D.

Sumithra J. Mandrekar, Ph.D.

Stephen S. Cha, M.S.

Alfred F. Furth, M.S.

Division of Biostatistics, Mayo Clinic, Rochester, MN, 55905

1. Introduction

The practice of dichotomizing continuous covariates is common in medical and epidemiological research for several reasons, both clinical and statistical. From a clinical point of view, binary covariates may be preferred for (1) offering a simple risk classification into "high" versus "low", (2) establishing eligibility criteria for prospective studies, (3) assisting in making treatment recommendations, (4) setting diagnostic criteria for disease, (5) estimating prognosis, and (6) imposing an assumed biological threshold [1-7]. From a statistical point of view, binary covariates may be preferred for (1) offering a simpler interpretation of common effect measures from statistical models such as odds ratios and relative risks, (2) avoiding the linearity assumption implicit in common statistical models for continuous covariates, (3) modeling a previously suspected or assumed threshold effect, and (4) making data summarization more efficient [1, 7-11].

Data dependent methods for dichotomizing continuous covariates, such as splits about some percentile (median, 25^{th}, 75^{th}), the mean, or some reputed clinically relevant threshold are arbitrary and may not be useful in assessing a variable's true prognostic value [1, 4, 6, 9, 12-14]. In contrast, outcome-based methods allow an "optimal" cutpoint to be estimated; the "optimal" cutpoint being defined as that threshold value of the continuous covariate distribution which, using statistical criteria, best separates low and high risk patients with respect to some outcome [2, 3, 5-7, 9, 14-17]. While methods utilizing statistical criteria allow for estimation of a "best" threshold value for a covariate, the inherent difficulties with searching for and utilizing an optimal cutpoint are well documented, namely the inflation of Type I error rates, a tendency to overestimate measures of effect, the potentially substantial loss of information when categorizing, and the inability to replicate the optimal cutpoint in subsequent studies [1-4, 6-10,

12-14, 17-19]. Several methods for calculating appropriate p-values and unbiased effect measures with optimal cutpoints have been documented in the statistical and medical literature, and will be discussed throughout this report.

The goal of this Technical Report is to consolidate the extant literature and describe in detail a unified strategy for finding optimal cutpoints with respect to binary and time-to-event outcomes, though analogous techniques for other types of outcomes also exist. Two in house SAS® macros which can perform the more salient procedures for identifying a cutpoint have been developed in conjunction with this Technical Report. It is important to note that the strategy described herewith is *most* appropriate when a threshold value truly exists. That is, we assume some binary split of the continuous covariate will create two relatively homogeneous groups with respect to a particular outcome [15, 20]. Under this assumption, a dichotomized version of the continuous covariate can be employed as the independent variable and the binary or time-to-event outcome as the dependent variable in what will be referred to as a *cutpoint model*. In a graphical sense, the *cutpoint model* can be thought of as a step function that adequately portrays the relationship between the continuous covariate and the outcome, where the risk of outcome remains at a constant level up to some cutpoint, then abruptly drops or rises vertically at the cutpoint to a new level of risk which remains constant throughout [6, 7, 11, 20, 21]. This is often an over-simplified and unrealistic portrayal of the true underlying model, but nevertheless may be preferable to a linear model or a more sophisticated model incorporating flexible functions for continuous covariates that may overfit idiosyncrasies in a particular data set and/or be difficult to interpret and communicate to investigators [1, 2, 11, 20]. Notwithstanding the above discourse, the *cutpoint model* is a reasonable alternative (and a valid

model) under any monotonic relationship between the continuous variable and outcome, a situation characterizing most covariate-outcome relationships [2, 11].

2. General Strategy

In this section a general strategy for finding optimal cutpoints is described, beginning with determining the appropriateness of a cutpoint model via (a) graphical diagnostic plots, followed by (b) estimation of a cutpoint. A description of the two SAS® macros, %cutpoint and %findcut, in the context of binary and time to event outcomes respectively is provided in the appendix.

2a. *Graphical diagnostic plots*

In the absence of any a priori clinical information regarding the prognostic relationship between a continuous covariate and outcome, the appropriateness of a cutpoint model must be determined empirically with graphical and numerical results [4, 6, 15, 16]. Several techniques have been introduced which allow flexible forms of a continuous covariate to be modeled, including smoothing splines, fractional polynomials, and non-parametric regression [21, 22]. Of these, smoothing splines have received the most attention, and have been incorporated into some of the standard statistical software packages. The results from models utilizing splines allow one to graph the continuous covariate as a function against the modeled outcome or some transformed version of the outcome such as odds ratios or hazard ratios from a logistic or Cox regression model respectively. A steep and definitive increase or decrease in the spline function near a threshold value which is relatively flat before and after the threshold provides evidence in favor of a *cutpoint model*. Grouped data plots and plots of Martingale residuals also serve as valuable graphical diagnostics for the appropriateness of a *cutpoint model* in a logistic or Cox

regression model setting respectively [6, 23]. These techniques will be described in Sections 3 & 4 respectively.

2b. Estimation of optimal cutpoint

Determining the existence of a threshold effect and estimating an optimal cutpoint for a single continuous covariate uses a series of two-sample tests for the multiple possible candidate dichotomizations of the continuous covariate. The maximum number of candidate cutpoints is k-1, where k is the number of unique values of the continuous covariate. Some have suggested excluding the outer 10-20% of the continuous covariate distribution to avoid having small numbers in one of the groups following dichotomization, thereby preventing substantial losses in statistical power [1, 5, 6, 16, 17]. The inner 80-90% of the distribution from which a cutpoint is chosen is referred to as the *selection interval*. For each candidate cutpoint within a specified selection interval, an appropriate two-sample test with concomitant test statistic and p-value (P_c) is determined. A *cutpoint model* may be appropriate if any P_c is less than or equal to some pre-specified allowable level of Type I error. The *optimal cutpoint* is often defined as that candidate cutpoint with the smallest P_c. This method for estimating a cutpoint is referred to as the *minimum p-value* approach, or alternatively the *maximum statistic* approach [1]. Other criteria for choosing an optimal cutpoint have been suggested, including maximum effect size and maximum precision of estimates, but have received less support [6, 24].

It is well recognized that Type I error rates for an optimal cutpoint found via the minimum p-value approach can be substantially inflated as a result of multiple comparisons [3, 4, 6, 10, 14, 18]. That is, the likelihood of finding a significant association at significance level α between a chosen cutpoint and outcome when in reality no relationship exists is likely to be much higher than α. This inflation rises as the number of candidate cutpoints examined

increases [4, 6, 10, 14]. In simulation studies Type I error rates have been found to be as high as 50% when examining 50 cutpoints [14]. Clearly the prognostic significance of a continuous covariate can be drastically overestimated with this dichotomization approach due to the series of statistical tests. Several alternative methods attempt to correct this problem: (1) significance level (α) adjustment, (2) p-value adjustment, and (3) cross-validation / split sample approach. The simplest and perhaps most popular method of α-level adjustment is the Bonferroni correction, which simply divides the desired pre-specified α-level (usually 0.05) by the number of candidate cutpoints examined [6, 10, 15]. Bonferroni correction assumes all statistical tests being conducted are independent of each other, and therefore is considered a conservative technique under non-independence of tests, in that the Bonferroni-adjusted α-level will be smaller than an adjusted α-level that accounts for the correlation between statistical tests [1, 6, 10, 20]. The technique's conservatism can be an attractive feature in determining the statistical significance of an optimal cutpoint, as only strong relationships between the dichotomized covariate and outcome would be deemed statistically significant.

Multiple authors have proposed p-value adjustment formulae, which use mathematical functions of observed p-values to estimate p-values adjusted for examining multiple candidate cutpoints [2, 5, 16, 17]. In the approach by Miller and Seigmund [17], the distribution of a maximally selected test statistic is derived, which is then used to obtain an adjusted p-value. An adjusted p-value of approximately 0.002 is equivalent to an unadjusted p-value of 0.05 when examining the inner 80% of the continuous covariate's distribution [1]. Furthermore, when examining the inner 90% of the distribution, an adjusted p-value of approximately 0.001 equates to an unadjusted p-value of 0.05 [1]. This approach requires the width of the selection interval to be specified.

The two-fold cross-validation technique provides a third method for determining the statistical significance of a cutpoint model [3, 7]. The steps involved in the technique are as follows:

(1) Randomly split all observations into two data sets of approximately equal size (call these Set I and Set II).

(2) Using Set I, find the cutpoint with the minimum p-value (C_1).

(3) Using Set II, find the cutpoint with the minimum p-value (C_2).

(4) Using C_1, group patients in Set II into High (above C_1, H) and Low (below C_1, L).

(5) Using C_2, group patients in Set I into High (above C_2, H) and Low (below C_2, L).

(6) Use the High and Low classifications to calculate the two-sample test statistic.

In simulation studies, Type I error rates from two-fold cross-validation were found to be approximately correct [3]. Two-fold cross-validation has the additional benefit of providing approximately unbiased estimates of effect size. Finally, the split sample approach can also be used to calculate p-values that are not affected by the multiple testing procedures. With this approach a data set is divided into a training set and a validation set. An optimal cutpoint is determined from the training set, and its significance and effect size determined from the validation set. While able to produce approximately correct p-values and effect sizes, this method does not use the entire data set in estimating an optimal cutpoint and often needs a large data set in which to split the data [3, 7]. In this report, we do not pursue the two fold cross validation and the split sample approaches.

2c. Summary of strategy and SAS Macro Content

In this Technical Report two SAS macros, %cutpoint and %findcut, are introduced which execute the aforementioned techniques for binary and time to event outcomes respectively. The

organization of each macro parallels the enumeration in the general strategy just described. Each macro begins with graphical diagnostic plots for determining the appropriateness of a *cutpoint* model, with grouped data plots and Martingale residual plots included in the %cutpoint and %findcut macro respectively. In the case of binary outcome data, within a user-specified selection interval, all candidate cutpoints are considered and corresponding unadjusted p-values are calculated. Furthermore, the corresponding adjusted p-values for each candidate cutpoint are calculated within each macro. See appendix I and II for more details on the macro parameters.

3. *Binary Outcome - Logistic Regression*

The analysis of a continuous covariate and a binary outcome lends itself to the logistic regression model. Logistic regression models the log odds of experiencing the binary event against any number of covariates as given below:

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \sum \beta_i x_i \ ,$$

where π is the probability of experiencing the outcome, x_i's are the covariates, and β_i's are the regression coefficients. The simple scenario of only a single continuous covariate is described here.

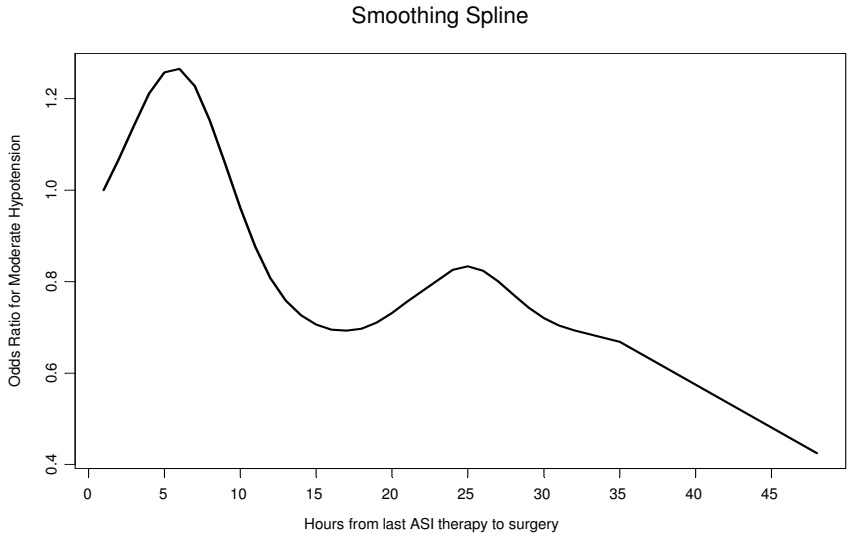
Throughout Section 3 the methodology for selecting an optimal cutpoint with a binary outcome is described in the context of a study examining the effects of angiotensin system inhibitor (ASI) discontinuation on hypotension during surgery with general anesthesia [25]. The binary outcome is the development of hypotension within 30 minutes following induction with an anesthetic agent. Hypotension is defined as a systolic blood pressure ≤ 85 mmHg. The continuous covariate is the time from last ASI treatment to anesthetic induction, measured in hours. All patients in this study were taking ASI therapy for high blood pressure prior to the index surgery, but the ASI therapy must be stopped prior to surgical induction. Prior studies

have suggested that continuing ASI therapy immediately up to surgery may lead to dangerously low blood pressure levels following anesthetic induction. The hypothesis was that longer intervals of time between last ASI therapy and surgery would lead to lower incidence of hypotension.

3a. Graphical diagnostic plots

Logistic regression models incorporating smoothing splines allow for a visual assessment of the functional relationship between the continuous covariate and outcome. A sharp increase or decrease in the spline function may suggest a cutpoint model is appropriate. At the very least a dichotomy with respect to outcome between low and high values of the covariate should be evident. Figure 1 depicts the functional relationship between time between last ASI therapy and surgery versus the odds for developing hypotension using a smoothing spline [Note: This was obtained using the S-plus spline function with 4 degrees of freedom].

Figure 1: Smoothing spline to illustrate the functional relationship between the covariate and outcome of interest

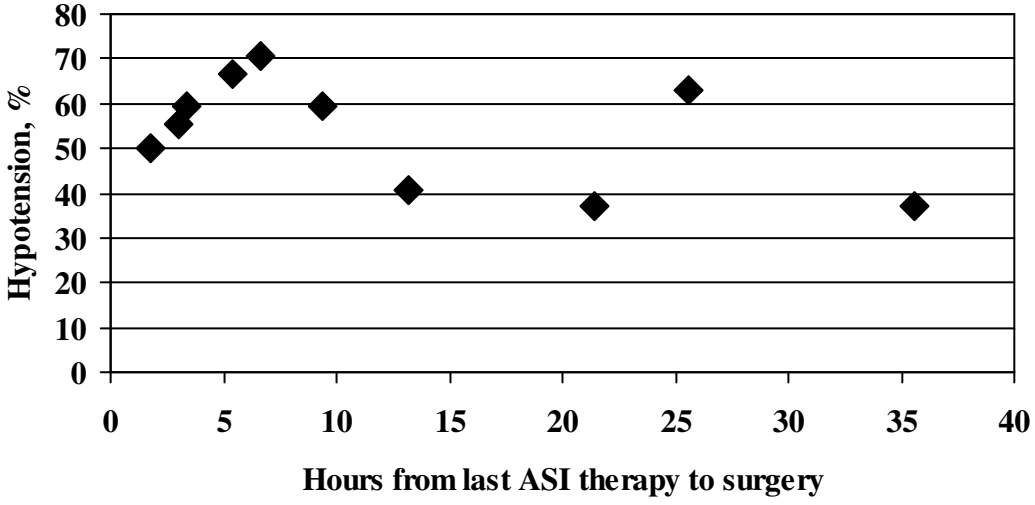


The vertical axis of Figure 1 indicates the odds ratios for hypotension with the smallest observed time between last ASAI therapy and surgery (1 hour) as the reference group for odds

ratio estimates. From this graph the appropriateness of a cutpoint model could be reasonably argued. A dichotomy in risk between low and high values of the covariate is evident. The two local maximums at approximately 5 and 25 hours cannot be explained clinically, and are likely aberrations resulting from a spline's tendency to overfit. Furthermore, a precipitous drop in the spline occurs beginning shortly after 5 hours and reaching a minimum shortly after 15 hours. One would suspect the optimal cutpoint to fall somewhere within these boundaries.

Grouped data plots are also a useful graphical diagnostic with a binary outcome [6]. These are created by grouping the continuous covariate into deciles (or some other quantile grouping) and then plotting the mean covariate value within each decile against the proportion experiencing the outcome in that decile (see Figure 2). For this data set, the grouped data plot gives a similar result as the spline.

Figure 2: Grouped data plot



3b. ***Estimation of optimal cutpoint***

For a binary outcome, examining a collection of candidate cutpoints can be reduced to a series of 2 by 2 tables,

	$X \le c$	$X > c$
Y=0	n_{11}	n_{12}
Y=1	n_{21}	n_{22}

where X is the continuous covariate to be dichotomized, c is a candidate cutpoint, Y is the binary outcome (0=no, 1=yes), and n_{11}, n_{12}, n_{21}, n_{22} are the respective cell counts. A p-value (P_c) can be calculated for each of the candidate cutpoints from a chi-square test: $P_c = P(\chi^2 > \chi_c^2)$. For the ASI therapy example, the ten candidate cutpoints, odds ratios (probability of developing or experiencing hypotension if hours from last ASI therapy to surgery is less than the candidate cutpoint compared to greater than the candidate cutpoint), and the corresponding unadjusted and adjusted p-values (using the approach of Miller and Seigmund, 1982) are shown in Table 1. The candidate cutpoints are ranked based on a total score obtained from the unadjusted p-value (highest p-value assigned the lowest score) and corresponding odds ratio estimate (lowest odds ratio estimate assigned the lowest score).

Table 1. Ten Candidate Cutpoints.

			P-value	
Hours from last ASI therapy to Surgery	Total Score	Odds Ratio (score)	Adjusted	Unadjusted (score)
9	16	1.782 (6)	0.339	0.020 (10)
10	13	1.767 (4)	0.360	0.021 (9)
29	13	2.488 (10)	0.741	0.069 (3)
27	12	2.137 (7)	0.595	0.047 (5)
28	12	2.151 (8)	0.705	0.063 (4)
8	11	1.754 (3)	0.376	0.022 (8)
30	11	2.463 (9)	0.879	0.099 (2)
11	9	1.733 (2)	0.433	0.028 (7)
12	7	1.724 (1)	0.461	0.031 (6)
26	6	1.770 (5)	0.923	0.112 (1)

In this example, a Bonferroni correction to the usual level of allowable Type I error (0.05) gives a corrected significance level of approximately 0.002 (dividing the Type I error rate by 27, the number of candidate cutpoints in the 90% selection interval). As stated earlier, Bonferroni adjustment is a conservative technique for judging the statistical significance of an optimal cutpoint, and in this case, none of the candidate cutpoints would be deemed statistically significant. Similarly, none of the candidate cutpoints for number of hours from last ASI therapy to surgery is significant based on the adjusted p-value approach.

4. Time-to-Event Outcome – Survival Analysis

An important difference between the outcome variables modeled via linear and logistic regression analyses and the time to event outcome is the fact that we may observe the outcome (survival, time to progression etc.) only partially. In other words, for those subjects who experience the event, we have the complete data or the actual time, whereas for subjects who do not experience the event or are lost to follow up, we only have the length of the follow-up, which is an incomplete observation as we do not have the actual event times. These incomplete observations are referred as censored observations, which can fall into 3 categories: right censored, left censored or interval censored. In addition, incomplete observations can also occur due to a selection process inherent in the study design, which is referred to as truncation in this setting [26, 27]. Survival Analysis is thus the analysis of such data that corresponds to a time from a well-defined starting point until the occurrence of a well-defined event of interest or a pre determined end of the observation period.

The fundamental building block of survival analyses is the cumulative distribution of the survival times, $S_T(t)$, and the hazard function, h(t), or the instantaneous failure rates. If T is the random variable denoting survival time, then $S_T(t) = P(T \geq t) = P$(an individual survives beyond

time t) = 1-P(T<t), h(t) = -d/dt log(S_T(t)), where 0<t<∞. Some of the outcome-oriented methods for cutpoint determination in a survival analysis setting are based on log rank, score, likelihood ratio and Wald statistics. Generally, the outcome-oriented methods are expected to have better statistical indicators than data-oriented methods [28]. In this report, we focus on the method proposed by Contal and O'Quigley [2], which is based on the log rank test statistic.

We use data from a multicenter trial of bone marrow transplant patients with a radiation-free conditioning regimen [29]. A total of 137 patients were classified into three disease groups: acute lymphoblastic leukemia (ALL, n = 38), acute myelocytic leukemia (AML) with low risk of first remission (n = 54), and AML with a high risk of second remission or untreated first relapse or second or greater relapse or never in remission (n = 45). Several potential risk factors were measured at the time of transplantation like recipient (patient) and donor sex, recipient and donor immune status, recipient and donor age (in years), waiting time (in months) from diagnosis to transplantation etc. However, for the purposes of illustration in this report, we only consider the following variables: patient's age, disease group, the outcome variable of interest, which is time to relapse or death (in months) along with a censoring indicator for relapse or death [27, 30].

4a. ***Graphical diagnostic plots***

The lowess smoothed plot of the martingale residuals is a graphical representation of an outcome-oriented approach to determine a cutpoint for the patient's age from the three disease groups. A stochastic process with a property that it's expected value at time t, given it's history at time s < t, is equal to it's value at time s, is called a martingale. Martingale residuals are used to determine the functional form of a covariate [see 23, 27 for derivation and discussion of the properties of martingale residuals]. PROC LOESS option in SAS® performs lowess smoothing with default smoothing parameter as 0.5 [26, 31]. There are several strategies that can be used to

select the smoothing parameter [31]. In our illustration we examine plots of the fitted residuals versus the predictor variable and choose the largest smoothing parameter that yields no clearly discernible trends in the fit residuals. Figures 3, 4, and 5 give the lowess smoothed residuals for the three disease groups: ALL, AML-Low, and AML-High respectively.

Figure 3: Plot of martingale residuals versus age and lowess smooth for ALL disease group

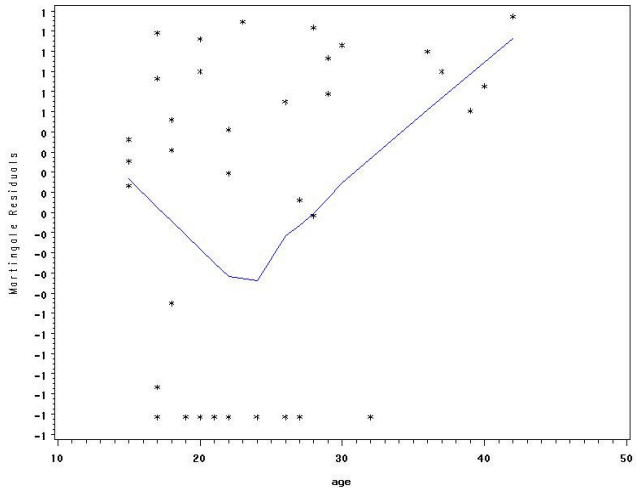


Figure 4: Plot of martingale residuals versus age and lowess smooth for AML-Low disease group

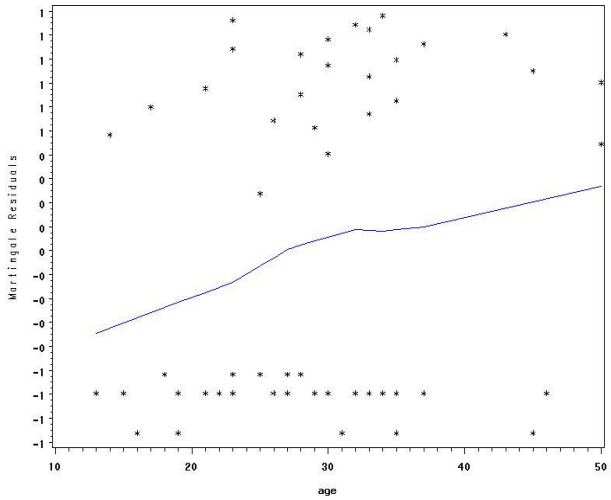
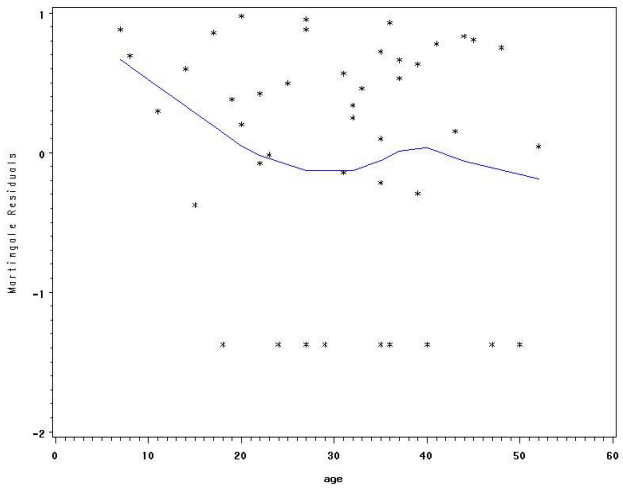


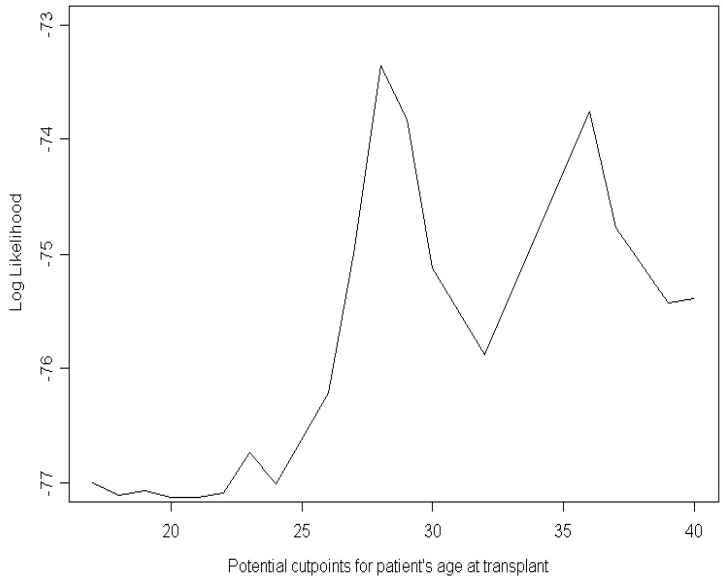
Figure 5: Plot of martingale residuals versus age and lowess smooth for AML-High disease group



The display of both the smooth fit and the individual residuals provides insight into the influence of specific individuals on the estimate of the functional form. Figure 3 suggests that treating age as linear is inappropriate for the ALL disease group. The smoothed curve is roughly zero up to about 24 years and increases linearly up to about 42 years. This suggests that patient's age can be coded as an indicator variable in the Cox proportional hazards model.

For distinct values of age, we create an indicator variable and then fit the Cox model with this new covariate to get the log-likelihood. The value of age that maximizes the log-likelihood gives the optimal cutpoint. For the ALL group, this occurs at 28 years as can be seen from Figure 6. However, in case of the AML-Low and AML-High groups, the lowess smooth values are nearly a straight line and support treating age as linear in the model (see Figures 4 and 5). Therefore, based on this approach, it is not appropriate to convert age into a categorical variable for the AML-Low and AML-High disease groups.

Figure 6: Plot of log-likelihood versus distinct patient ages at transplant for ALL disease group.



4b. ***Estimation of optimal cutpoint***

In this section, we focus on the method proposed by Contal and O'Quigley [2], which is based on the log rank test statistic. Let R be the risk factor of interest measured as a continuous variable and T be the outcome variable. In case of survival analysis, the outcome of interest T, is oftentimes time to death but it can also be time to some other event of interest. The population is divided into two groups based on the cutpoint: subjects with the value of the risk factor less than or equal to the value of the cutpoint and subjects with the value of the risk factor greater than the cutpoint. Let $t_{(1)} < t_{(2)} <.....< t_{(k)}$ be the ordered observed event times of the outcome variable T. Let C be the set of K distinct values of the continuous covariate R. Then, based on one hypothetical cutpoint from C, let d_i be the number of events at time $t_{(i)}$, r_i be the number of subjects at risk prior to time $t_{(i)}$ and d_i^+ and r_i^+ be the number of events at time $t_{(i)}$ in group R > C and number of subjects at risk just prior to $t_{(i)}$ in the group R > C. Similarly, d_i^- and r_i^- be the

number of events at time $t_{(i)}$ in group $R \leq C$ and number of subjects at risk just prior to $t_{(i)}$ in the group $R \leq C$. Thus, the log rank statistic for some fixed C is given by:

$$\text{Log Rank Statistic} = L_k(t) = \sum_{i=1}^{k} \left(d_i^+ - d_i \frac{r_i^+}{r_i} \right)$$

The optimal cutpoint is that value of C, C_k that maximizes the absolute value of Lk (t). C_k therefore gives the value of the continuous covariate that gives the maximum difference between the subjects in the two groups defined by the cutpoint. In order to test the significance of the obtained cutpoint, following test statistic is proposed:

$$\text{Test Statistic} = q = \frac{1}{s\sqrt{k-1}} \max \left| L_k(t) \right|$$

where, $s^2 = \frac{1}{(k-1)} \sum_{j=1}^{k} a_j^2$, and a_j's are the scores associated with the j^{th} death, given by

$1 - \sum_{i=1}^{j} \frac{1}{k - i + 1}$. Such a maximization of the statistic enables the estimation and evaluation of the significance of the cutpoint and is adjusted for the bias created by the fact that the optimal cutpoint C_k is chosen such that it gives the maximum separation between the two groups [2]. For $q > 1$, the p-value is approximately given by $2e^{-2q^2}$ and for $q \leq 1$, the p-value is at least 0.33. Thus, this procedure considers all possible values of the continuous covariate as potential cutpoints.

We now present the results for categorizing patients into high or low risk groups for disease free survival based on the patient's age at transplantation for the three groups and also assess the significance of the cutpoint [2]. In the ALL group, there are 20 distinct ages, any of which can be a potential cut point. There are 23 distinct times when death or relapse occurs, which gives $s^2 = 0.8757$. The maximum value of $|L_k|$ occurs at age 28 with $q = 1.2946$ and p-

value of 0.07 (see Table 2). This suggests that the cutpoint obtained is significant, i.e., age is related to time to disease free survival for ALL group (note: 10% level of significance is used due to the small sample size).

Table 2: Results for the ALL disease group.

Obs	Distinct Ages	L_k	$\|L_k\|$	q
1	15	0.0000	0.0000	0.0000
2	17	-0.7503	0.7503	0.1709
3	18	-0.4232	0.4232	0.0964
4	19	-0.7427	0.7427	0.1692
5	20	0.2725	0.2725	0.0621
6	21	-0.2736	0.2736	0.0623
7	22	0.7416	0.7416	0.1690
8	23	2.1637	2.1637	0.4930
9	24	1.2171	1.2171	0.2773
10	26	3.2475	3.2475	0.7399
11	27	4.7287	4.7287	1.0773
12	**28**	**5.6822**	**5.6822**	**1.2946**
13	29	4.7785	4.7785	1.0887
14	30	3.4222	3.4222	0.7797
15	32	2.5916	2.5916	0.5904
16	36	3.6068	3.6068	0.8217
17	37	2.8075	2.8075	0.6396
18	39	2.1071	2.1071	0.4801
19	40	1.6014	1.6014	0.3648
20	42	0.9737	0.9737	0.2218

In the case of AML-Low group, there are 26 distinct ages, any of which can be a potential cutpoint. There are 25 distinct times when death or relapse occurred which gives s^2 = 0.8827. The maximum value of $|L_k|$ occurs at age 28 with q= 0.983. However, the high p-value (> 0.33) suggests that the cutpoint obtained is not significant. This is also the case for AML-High risk group (31 distinct ages, 33 distinct times when death or relapse occurred, s^2 = 0.9035, q = 0.1464, p-value $\geq$ 0.33). Note that in this example, both the graphical and the estimation approach gives the same cutpoint for the ALL disease group, however, this need not be true in

general. In situations when the estimated cutpoint is close to boundaries, one should carefully examine the reasons as the cutpoint obtained may be real or may be due to the presence of outliers.

5. Discussion

Given the widespread use of categorizing a continuous covariate, there is very little attention given to this topic in statistical and epidemiological textbooks and in the literature. We have only focused on the dichotomization of a continuous covariate with the assumption that such a dichotomization is possible from biological point of view, however, in reality, more than one cutpoint may exist. Our current work provides an insight into some of the outcome-oriented cutpoint determination methods as well as SAS® macros that provide p-values adjusted for examining multiple candidate cutpoints.

Ideally this cutpoint search has to be done within the framework of a multiple regression model to eliminate the potential influence of other prognostic factors on the cutpoint. As stated before, one has to be also aware of potential confounding that might arise from categorization and using open-ended categories [32]. The obtained cutpoint(s) may differ across studies depending on several factors including which data or outcome-oriented approach is used and therefore the results may not be comparable. Lastly, there is always the possibility of loss in information from categorizing a continuous covariate, possible loss of power to detect actual significance and can sometimes lead to biased estimates in regression settings, all of which need to be sufficiently addressed [33, 34].

AUTHOR CONTRIBUTIONS

Brent Williams and Alfred Furth contributed to the work on cutpoint determination methods for a continuous covariate with a binary (logistic) outcome; Jay Mandrekar, Sumithra Mandrekar, and Steve Cha contributed to the work on cutpoint determination methods for a continuous covariate with time to event (survival) outcome.

References

1. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. Journal of the National Cancer Institute 1994; 86: 829-835.
2. Contal, C., O'Quigley, J. "An application of changepoint methods in studying the effect of age on survival in breast cancer," Computational Statistics and Data Analysis, 1999; 30, 253 - 270.
3. Faraggi D, Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. Statist Med 1996; 15: 2203-2213.
4. Hilsenbeck SG, Clark GM. Practical p-value adjustment for optimally selected cutpoints. Statistics in Medicine 1996; 15: 103-112.
5. Lausen B, Schumacher M. Maximally selected rank statistics. Biometrics 1992; 48: 73-85.
6. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. Statistics in Medicine 2000; 19: 113-132.
7. Mazumdar M, Smith A, Bacik J. Methods for categorizing a prognostic variable in a multivariable setting. Statistics in Medicine 2003; 22: 559-571.
8. Cumsille F, Bangdiwala SI, Sen PK, et al. Effect of dichotomizing a continuous variable on the model structure in multiple linear regression models. Commun Statist – Theory Meth 2000; 29: 643-654.
9. Heinzl H, Tempfer C. A cautionary note on segmenting a cyclical covariate by minimum p-value search. Computation Statistics and Data Analysis 2001; 35: 451-461.

10. Liquet B, Commenges D. Correction of the p-value after multiple coding of an explanatory variable in logistic regression. Statistics in Medicine 2001; 20: 2815-2826.
11. Royston P, Sauerbrei W, Altman DG. Modeling the effects of continuous risk factors. Journal of Clinical Epidemiology 2000; 53: 219-222.
12. Altman DG. Categorising continuous variables. Br J Cancer 1991; 64: 975.
13. Courdi A, Hery M, Chauvel P, et al. Prognostic value of continuous variables in breast cancer and head and neck cancer. Dependence on the cut-off level. Br J Cancer 1988; 58: 88-90.
14. Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out? Breast Cancer Research and Treatment 1992; 22: 197-206.
15. Abdolell M, LeBlanc M, Stephens D, et al. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Statist Med 2002; 21: 3395-3409.
16. Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. Computational Statistics and Data Analysis 1996; 21: 307-326.
17. Miller R, Siegmund D. Maximally selected chi square statistics. Biometrics 1982; 38: 1011-1016.
18. Hollander N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an `optimal' cutpoint. Statistics in Medicine, 2004; 23:170-713.
19. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. Epidemiology 1992; 3: 434-440.

20. Schulgen, G., Lausen, B., Olsen, J. H., and Schumacher, M. "Outcome-oriented cutpoints in analysis of quantitative exposures," American Journal of Epidemiology, 1994; 140, 172 - 184.

21. Greenland S. Avoiding power loss associated with categorization and ordinal scores in does-response and trend analysis. Epidemiology 1995; 6: 450-454.

22. Harrell, F.E. Jr.. Regression modelling strategies with applications to linear models, logistic regression, and survival analysis, 2001; Springer-Verlag, New York.

23. Therneau T. M., Grambsch, P.M., Fleming, T. R. "Martingale-based residuals for survival models," Biometrika, 1990; 77, 147 - 160.

24. Magder LS, Fix AD. Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. J Clin Epidemiol 2003; 56: 956-962.

25. Comfere T, Sprung J, Kumar MM, Draper M, Wilson DP, Williams BA, Danielson DR, Liedl L, Warner DO. Angiotensin system inhibitors in a general surgical population. Anesth Analg. 2005;100(3):636-44.

26. Hosmer, D. W. Jr., Lemeshow, S. Applied Survival Analysis: Regression Modeling of Time to Event Data, 1999; New York: Wiley.

27. Klein, J. P., and Moeschberger, M. L. Survival Analysis: Techniques for Censored and Truncated Data, 1997; New York: Springer-Verlag.

28. Kuo, Y. "Statistical methods for determining single or multiple cutpoints of risk factors in survival data analysis," Dissertation, Division of Biometrics and Epidemiology, School of Public Health, The Ohio State University, 1997.

29. Copelan, E. A., Thompson, J. M., Crilley, P., Szer, J., Klein, J. P., Kapoor, N., Avalos, B. R., Cunningham, I., Atkinson, K., Downs, K., Harmon, G. S., Daly, M. B., Brodsky, I.,

Bulova, S. I., Tutschka, P. J. “Treatment for Acute Myelocytic Leukemia with Allogenic Bone Marrow Transplantation Following Preparation with Bu/Cy,” Blood, 1991; 78, 838 - 843.

30. Mandrekar JN , Mandrekar SJ , Cha SS . Cutpoint determination methods in survival analysis using SAS®. Proceedings of the 28th SAS Users Group International Conference (SUGI), 2003; 261-28.
31. Cohen, R. A. “An introduction to PROC LOESS for Local Regression,” Proceedings of the 24th SAS Users Group International Conference (SUGI), 1999; Paper 273, 1584 - 1592.
32. Rothman, K. J.,Greenland, S. Modern Epidemiology, 1992; 2nd Edition, Philadelphia: Lippincott-Raven.
33. Selvin, S. (1987), “Two issues concerning the analysis of grouped data,” European Journal of Epidemiology, 3, 284 - 287.
34. Altman, D. G. (1998), “Categorizing continuous variables,” in Armitage, P. and Colton, T. (eds), Encyclopedia of Biostatistics, Chichester: John Wiley, 563 - 567.

Appendix I

%cutpoint(dvar=, endpoint=, data=, trunc=, type=, range=, fe=, plot=plot, ngroups=, padjust=, zoom=);

This macro finds a cutpoint for a continuous variable when the outcome of interest is binary

Parameters:
dvar = continuous variable to dichotomize
endpoint = endpoint in question
data = dataset
trunc = type of truncation of continuous variable
round - normal rounding
int - moves to the integer in the direction of zero on the number line
floor - moves to the next left integer
ceil - moves to the next right integer
type = type of iteration
integers - iterates to next integer of cont. var.
tenths - iterates to the next tenth
hundredths - iterates to the next hundredth
range = range of continuous variable
fifty - inner 50% of cont. var. used for cutpoints
eighty - inner 80% of cont. var. used for cutpoints
ninety - inner 90% of cont. var. used for cutpoints
fe = perform fisher's exact test when expected cell counts are less than 5
on - turns fe on
off - turns fe off
plot = type of plot
plot: regular output window plots
gplot: gplot output
ngroups = number of groups to spilt the continuous
variable into (any integer > 1 and < n)
padjust - p-value adjusting technique
miller -
zoom - zoom into the minimum p-value plot
yes - zooms into the lower half of the p-values
no - no zooming

Sample code used for the example in section 3.0:

```
libname dat v6 '~furth/consult/s101370/data';
%cutpoint(dvar=aceilst2, endpoint=mhypo30, data=dat.anal, trunc=round, type=integers,
range=ninety, fe=on, plot=gplot, ngroups=10, padjust=miller, zoom=no);
```

Appendix II

%findcut (ds= , time= , stat= , cutvar= , maxtime=, printop=, residop=, plottype=);

This macro finds a cutpoint for a continuous variable with a time to event outcome.

Parameters:
ds = name of the dataset
time = variable containing the time to event information
stat = the status or event indicator, for example, 1=event or 0=no event
cutvar = the continuous variable for which the cutpoint is to be determined

Optional Parameters:
maxtime = Time point (in days) to get % survival, default is 182.5 days (i.e. 6 months)
printop = basic summary statistics. 0 is default.
1=proc univariate on a continuous cutvar
2=proc freq on a ordinal cutvar
0=No print
residop = Requests martingale residual plot.
0=No Martingale residual plot
1=plot Martingale residual plot
plottype = Requests the minimum p-value and the maximum hazard ratio plots. 1 is default
0 = No plot
1=prints to the greenbar printer
2=prints to the Unix laser printer
(Note: This option may not be relevant for users outside of Mayo)