New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

Dangers of Using 'Optimal' Cutpoints in the Evaluation of Cyclical Prognostic Factors

Harald Heinzl[1]

Abstract

It is common strategy in medical research to categorize a continuous covariable before evaluating its prognostic impact on clinical outcome. In most cases the covariable is divided into just two groups. The chosen cutpoint is either a value already published in other studies, or a certain sample quantile like the median, or a so-called 'optimal cutpoint', that is the value which corresponds to the most significant relation with outcome. Because the multiple testing problem is often ignored, the term 'optimal' is misleading in this context. Altman et al. (1994) suggest that the method be called the 'minimum P-value approach' instead, and present simulation and asymptotic results of the inflation of the type I error rate.

Recently the influence of menstrual status at the time of surgery on the prognosis of women suffering from breast cancer was discussed in the medical literature. Although the paper which triggered the discussion, reported a high relative risk for death in patients who underwent breast cancer surgery during the perimenstrual period, almost all of the subsequently published work could not confirm this result in retrospective studies.

The menstrual status at the time of breast cancer surgery is a cyclical covariable. Its splitting into two segments is a similar strategy of analysis like the categorization of a continuous covariable. In the case that this splitting is based on a minimum P-value search, the problem of multiple testing has to be taken into account, too. Following Altman et al. (1994), a simulation study was performed to gain some insight into the relation of the actual versus nominal type I error rate with regard to the breast cancer example.

[1] Department of Medical Computer Sciences, University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.

1 Introduction

It was suggested by Hrushesky et al. (1989) that the menstrual status at the time of surgery could be a potential prognostic factor for survival of premenopausal women suffering from breast cancer. In particular, surgery during the perimenstrual period (days 1-6 and 21-36 after the last menstrual period) was found to be more hazardous than during the mid-cycle (days 7 to 20). This result caught the attention of oncologists and surgeons, since the suggested new risk factor can be determined and controlled in an easy and cheap way. Patients would have better chances for survival just by adequately timing their surgery.

Unfortunately this finding could not be repeated in other retrospectively reviewed data sets, see references in McGuire (1991) or Tempfer et al. (1996). Taking into account the probable publication bias, we can reasonably assume that there are even more negative results arround, which editors of medical journals have refused to publish due to lack of power. Somewhat simplifying, we could conclude that Hrushesky et al. (1989) was a false positive result, which we expect to occur under the null hypothesis of "no menstrual cycle effect" on an average in one out of twenty cases just by chance, when using a significance level of five percent.

In the meantime an additional study suggested that time of surgery in days 3 to 12 of the menstrual cycle leads to a poorer prognosis within premenopausal breast cancer patients. In addition, several pathophysiological mechanisms which could possibly account for this findings have been discussed. Again, this could not be confirmed by other studies, although yet another team of physicians claimed that surgery done on days 7 to 14 from the start of the last menstrual period should be avoided within a certain subgroup of the patients, that is, patients with metastatic disease in the axillary lymph nodes (usually termed "node-positive").

All these rather contradictory findings in mind, McGuire (1991) mentioned two important issues. First, he pointed out the retrospective nature of all the analyses. Unbalanced patients characteristics, confounding treatment effects, follow-up bias, and so forth, can lead to inconsistent, biased and spurious results. Secondly, he posed the question for the validity of the statistical techniques used to analyse the data sets in question. In particular, he addressed the problem of so-called "optimal cutpoint" analysis. That is, trying various menstrual cycle segmentations in search for significance. If such a procedure is applied without properly dealing with the inevitably arising multiple testing problem, the actual false positive error rate will be much higher than the chosen significance level reported in the study.

This paper will examine McGuire's (1991) statistical concerns in greater detail. In Section 2 the multiple testing problem of "optimal cutpoint" search for a

simple continuous covariate will be studied in the context of survival analysis according to Altman et al. (1994). A simulation study will by designed to adapt for the fact that menstrual cycle is a cyclical covariate. The simulation results will be described in Section 3. The outcome of an "optimal cutpoint" search in a real data set of 149 patients will be adjusted according to the findings of this simulation study. In Section 4, further aspects of the problem will be mentioned and discussed.

2 Methods

The main aim of a considerable number of oncological research papers is to investigate the importance of potential prognostic factors on a failure time outcome variable like overall survival or disease free survival. If such a factor is measured on a continuous scale, then in most cases it will be categorized into two or more groups. Categorization enables researchers to avoid strong assumptions about the relation between covariable and outcome variable, but at the expense of throwing away information, Altman et al. (1994). The information loss is naturally greatest with only two groups, but this approach is most common. In the following, the validity of this analysis technique will not be questioned anymore until the discussion in Section 4.

The most wide-spread methods to choose a cutpoint are: (i) Use a value already published in other studies, (ii) use a certain sample quantile like the median, or (iii) use a so-called 'optimal cutpoint', that is the value which corresponds to the most significant relation with outcome. The latter has become quite popular among clinical researchers. This has to be considered unfortunate, since the accompanying multiple testing problem will be ignored in general. Due to the term 'optimal', this questionable method is often considered superior by non-statisticians. Altman et al. (1994) suggest that the method be called the 'minimum P-value approach' instead.

The minimum P-value approach requires the choice of a selection interval. The selection interval is characterized by the proportion ε of smallest and largest values of the continuous covariate that are not considered as potential cutpoints. The cutpoint is varied systematically within the selection interval, a P-value is computed for each cutpoint, and the cutpoint with the smallest P-value is chosen eventually. The different test statistics involved are not independent so that the well-known Bonferroni-Holm correction (Holm, 1979) is not adequate to deal with the multiple testing problem. The inflation of the type I error rate for the logrank test has been studied using theoretical arguments (Lausen and Schumacher, 1992) and simulation studies (Hilsenbeck et al., 1992, Altman et al., 1994). The logrank test has been studied since it is standard choice for testing group differences in survival times outcome variables.

Lausen and Schumacher (1992) showed that the maximum of the absolute value of the standardized logrank statistic converges in distribution to the supremum of the absolute value of a standardized Brownian bridge. Their theoretical considerations and some earlier results of Miller and Siegmund (1982) allow a correction, valid for large sample sizes, of the minimal P-value to allow for the multiple testing. If $P_{\min}$ denotes the minimum P-value of the logrank statistic, the corrected P-value, P_{cor}, can be obtained as follows:

$$P_{\text{cor}} = \varphi(z)\left(z - \frac{1}{z}\right)\ln\left(\frac{(1-\varepsilon)^2}{\varepsilon^2}\right) + \frac{4\varphi(z)}{z},$$

where φ denotes the standard normal density and z is the $[1-(P_{\min}/2)]$-quantile of the standard normal distribution. According to Altman et al. (1994), there are simpler approximations available in the case of small minimum P-values, that is, $0.0001 < P_{\min} < 0.1$, specifically,

$$P_{\text{cor}} \approx -1.63\, P_{\min}(1 + 2.35 \ln P_{\min}),\ \textit{for}\ \varepsilon = 0.10$$
$$P_{\text{cor}} \approx -3.13\, P_{\min}(1 + 1.65 \ln P_{\min}),\ \textit{for}\ \varepsilon = 0.05$$

These formulas and the simulation results show that there is an increase in the actual false-positive error rate when the selection interval is increased, that is, when the proportion ε is decreased. The simulations results further show that there is hardly any dependence on sample size, see Altman et al. (1994).

The time from the start of the last menstrual period is a cyclical covariable, and the above results can not be applied. A simulation study was designed to overcome this problem. At first, we have to adapt the definition for a selection interval. When categorizing a cyclical covariable by forming two groups, then two cutpoints have to be chosen. The first cutpoint marks both the beginning of the first segment and the end of the second segment on the circle, whereas the second cutpoints marks the end of the first and the beginning of the second segment, respectively. It follows that the selection interval on a circle is characterized by the proportion $\varepsilon_{\text{circle}}$ of the minimum segment length allowed.

We used randomly generated exponentially distributed survival data, assuming all patients have a constant hazard of failure over time. A constant menstrual cycle length of 28 days was assumed and a menstrual cycle value between 1 and 28 days was randomly assigned to every survival time. The sample size (n=140, n=280, n=1400), the amount of censoring (33% and 67%), and the minimum selection interval (7 days and 14 days) were varied. Subsequently, 2000 simulated samples were generated for each of these 12 different scenarios.

If we use a minimum length of the selection interval of 14 days, then there will be 14 non-redundant partitions of the 28 days long menstrual cycle. That is, we

can test partition (1-14) vs. (15-28), (2-15) vs. (16-1) and so on until (14-27) vs. (28-13). Note that the partitions (15-28) vs. (1-14), (16-1) vs. (2-15) until (28-13) vs. (14-27) are redundant here. In the case we use a minimum segment length of 7 days, there are 210 different partitions possible. That is, we can test all 28 (7:21)-days partitions, then all 28 (8:20)-days partitions and so on until all 14 non-redundant (14:14)-days partitions.

All calculations were done by using the SAS statistical software system (SAS Institute Inc., Cary, NC, USA, 1990).

3 Results

The results of the simulation study to determine the amount of type I error in a multiple testing situation, designed as an analogue to the menstrual cycle, are shown in Tables 1 and 2. Table 1 shows the amount of type I error applying a minimum selection interval of 14 days. In Table 2 this interval was decreased to 7 days.

Table 1: Effect of the minimum P-value approach on the false-positive error rate for a minimum selection interval of 14 days. The nominal significance levels shown are 1%, 5%, and 10%. The results are based on 2000 simulated samples each.

Sample size	Percentage censored	Proportion of false-positive results observed		
		Nominal $\alpha = 0.01$	Nominal $\alpha = 0.05$	Nominal $\alpha = 0.10$
140	**33%**	0.069	0.27	0.46
140	**67%**	0.073	0.29	0.47
280	**33%**	0.073	0.27	0.49
280	**67%**	0.081	0.28	0.48
1400	**33%**	0.072	0.27	0.45
1400	**67%**	0.074	0.27	0.46

We further used the results of the simulation study to correct for the multiple testing of a minimum P-value search in an actual data set. One hundred and forty-nine Austrian patients suffering from breast cancer were included in the study. The

median follow-up time was 46.4 months. During the observation period, 50 patients showed recurrence of disease. Patients with menstrual status at the time of surgery greater then 28 days were given a value of 28 days.

Table 2: Effect of the minimum P-value approach on the false-positive error rate for a minimum selection interval of 7 days. The nominal significance levels shown are 1%, 5%, and 10%. The results are based on 2000 simulated samples each.

Sample size	**Percentage censored**	**Proportion of false-positive results observed**		
		Nominal $\alpha = 0.01$	Nominal $\alpha = 0.05$	Nominal $\alpha = 0.10$
140	**33%**	0.25	0.63	0.83
140	**67%**	0.27	0.63	0.84
280	**33%**	0.25	0.63	0.84
280	**67%**	0.24	0.63	0.83
1400	**33%**	0.23	0.60	0.81
1400	**67%**	0.23	0.62	0.82

To evaluate the prognostic value of menstrual status at the time of surgery on disease-free survival, a minimum P-value search with a minimum segment length of 7 days was applied. That is, 210 logrank tests were performed. The “best“ result found was a P-value of 0.011 with a corresponding bipartition of the menstrual cycle of (14-21) vs. (22-13), with a higher risk for recurrence of disease in segment (14-21). Note that by coincidence this uncorrected result would have closed the circle of suggested surgery times to avoid, see Section 1.

The distribution of 2000 minimum P-values found by simulation (sample size $n = 140$, 67% censoring, minimum segment length of 7 days) was used to obtain an approximative corrected P-value of 0.28 for the minimum P-value of 0.011 found in the actual data set, see Tempfer et al. (1996) and Haeusler et al. (1996).

4 Discussion

The results of Tables 1 and 2 are impressive. If a nominal significance level of 5% and a minimum segment length of 14 days is chosen, then approximately one quarter of all results will be false-positive under the null hypothesis of no

menstrual cycle effect. This proportion increases to approximately six out of ten, if the minimum segment length is decreased to 7 days. If observed P-values are in the interval [0.05, 0.10], then physicians will tend to call such a result a "trend". Having the results of Table II in mind, we have to consider all those researchers "unlucky", who are not able to discover at least some sort of "trend" in their data set by applying the uncorrected minimum P-value approach.

Tables 1 and 2 further show that the inflation of the type I error rate does not depend on the sample size. Also, no dependence on the censoring percentage could be detected. Besides that, it has been shown in Section 3 by way of example that the results of the simulation study can be easily used to correct the minimum P-values found in actual data sets.

In conclusion, there are some further points which should be mentioned:

- Are we measuring the menstrual status in a correct fashion? Instead of just counting the days since the last menstrual period, we could consider the fraction of menstrual status divided by the woman's usual menstrual cycle length to be a more accurate measure of the phenomenon. We could even think of using clinical laboratory methods to determine the actual menstrual status as accurately as possible.
- As in Lausen and Schumacher (1992) for a simple continuous covariate, theoretical considerations could be carried out to find an asymptotic formula to correct the minimum P-value of the logrank statistic for a cyclical covariate.
- If we use a cyclical covariate categorized by a minimum P-value search as the predictor in a Cox model, what is the effect on the corresponding regression coefficient? It seems to be quite obvious that we are overestimating it, but to which extent? A starting point for an investigation of this question could be Lausen and Schumacher (1995).
- The estimated cutpoint location of a minimum P-value search may be biased caused by sample size effects. Abel et al. (1984), and Abel and Berger (1986) suggested a simple method to overcome this problem in the case of an ordinary continuous covariate. Further research could be done to adapt their approach for a cyclical covariate.
- Besides all the popularity of dichotomizing continuous covariables by minimum P-value search in the medical literature, there are concerns among statisticians that, without biological indications of the actual existence of a cutpoint, the application of such an approach has to be considered methodologically inferior and should therefore be avoided. It seems to be desirable for clinical investigators to report statistical analyses with the covariables treated as continuous variables, applying smoothing or related techniques to explore the relationship with the clinical outcome. Confidence

bands should be added. These suggestions are also valid for cyclical covariables.

- When we consider the application of a cutpoint search within a continuous covariable to be a reasonable task, we should always take into account the possibility that more than just one cutpoint exists in reality. Furthermore, we should consider eliminating the potential influence of other prognostic factors on our search result, that is, we should perform the cutpoint search within the framework of a multiple regression model. Such a procedure can be seen to be equivalent to the knot placement problem with regression splines, see Heinzl (1994). Again, these suggestions can be adapted for cyclical covariables also.
- All the clinical studies mentioned have been of a retrospective nature. As already mentioned in Section 1, potential sources of bias of such studies are numerous. A valid way to overcome this problem is the following. If there are enough biological indications and potential explanations that there may be a relationship between menstrual status at the time of surgery and survival, then a randomized prospective study should be carried out. McGuire (1991) puts it in more drastic words: "In summary, low tide or high tide, I don't know, but let's find out." Until now the author is aware of only one randomized prospective study in progress, where the menstrual status at the time of surgery has been added to the list of potential prognostic factors to examine (Myles, 1996).

Finally, it should never be forgotten that behind oncological data there is the enormous distress of the patients and their relatives. To perform clinical studies in an appropriate and responsible way is a challenge for both scientific *and* ethical reasons.

References

[1] Abel, U., Berger, J., and Wiebelt, H. (1984): CRITLEVEL: An Exploratory Procedure for the Evaluation of Quantitative Prognostic Factors. *Methods of Information in Medicine*, **23**, 154-156.

[2] Abel, U. and Berger, J. (1986): Welche Fallzahlen erfordert die Methode CRITLEVEL? Ergebnisse einer Simulationsstudie. *EDV in Medizin und Biologie,* **17**, 9-11.

[3] Altman, D.G., Lausen, B., Sauerbrei, W., and Schumacher M. (1994): Dangers of Using 'Optimal' Cutpoints in the Evaluation of Prognostic Factors. *Journal of the National Cancer Institute*, **86**, 829-835.

[4] Haeusler, G., Tempfer, C., Kainz, C., and Heinzl H. (1996): Menstrual phase and timing of breast cancer surgery: statistical aspects. Letter to Editor. *British Journal of Cancer,* **74**, 1851-1852.

[5] Heinzl, H. (1994): Methoden zur Kategorisierung von stetigen Kovariablen in Regressionsmodellen (insbesondere bei medizin-statistischen Fragestellungen). *Dissertation an der Sozial- und Wirtschaft-wissenschaftlichen Fakultät der Universität Wien*.

[6] Hilsenbeck, S.G., Clark, G.M., and McGuire, W.L. (1992): Why do so many prognostic factors fail to pan out? *Breast Cancer Research and Treatment,* **22**, 197-206.

[7] Holm, St. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* **6**, 65-70.

[8] Hrushesky, W.J., Bluming, A.Z., Gruber, S.A., and Sothern R.B. (1989): Menstrual influence on surgical cure of breast cancer. *The Lancet,* **2**, 949-952.

[9] Lausen, B. and Schumacher, M. (1992): Maximally Selected Rank Statistics. *Biometrics,* **48**, 73-85.

[10] Lausen, B. and Schumacher, M. (1995): Evaluating the Effect of Optimized Cutoff Values in the Assessment of Prognostic Factors. *Technical Report No. 7 of the Freiburger Zentrum für Datenanalyse und Modellbildung,* Albert-Ludwig-Universität, Freiburg im Breisgau, 33 pages.

[11] McGuire, W.L. (1991): The Optimal Timing of Mastectomy: Low Tide or High Tide? Editorials, *Annals of Internal Medicine*, **115**, 401-403.

[12] Miller, R. and Siegmund, D. (1982): Maximally Selected Chi Square Statistics. *Biometrics*, **38**, 1011-1016.

[13] Myles, J.D., El Kum, N.B., and Levine, M. (1996): A method for modelling circadian rhythms in failure time data with application to predicting the best time for surgery in pre-menopausal breast cancer. The XVIIIth International Biometric Conference (IBC 96), Amsterdam, The Netherlands. *Contributed Papers*, **55**.

[14] Tempfer, C., Haeusler, G., Heinzl, H., Kolb, R., Hanzal, E., and Kainz, Ch. (1996): Menstrual phase and breast cancer surgery: influence on clinical outcome or pitfall of statistical analysis? *Cancer Letters,* **110**, 145-148.