Breast Cancer Research and Treatment **22**: 197-206, 1992.

Why do so many prognostic factors fail to pan out?

Susan Galloway Hilsenbeck, Gary M Clark, William L McGuire[1]
Division of Medical Oncology, Department of Medicine, University of Texas Health Science Center at San Antonio, TX, USA

Key words: data exploration, multiple cutpoints, prognostic factors, validation datasets

Summary

Although there can be many reasons that one study fails to confirm the results of another, the consequences of data exploration and the potential for spuriously significant results are often overlooked. A series of simulation experiments were designed to mimic the characteristics of relapse-free survival data that might be encountered in a prognostic factor study of node-negative breast cancer patients. Each simulated dataset of 500 or 250 cases was divided into a training set, used to select the "best" prognostic factor cutpoint, and a validation set, used to confirm the cutpoint. Testing multiple cutpoints markedly increased the risk of making a Type I error. The power to detect even small true differences was substantial, and increased as the number of cutpoints increased. Regardless of the number of cutpoints tested on the training sets, the Type I error rate on an independent validation data set was quite stable and the power of the validation set to detect true differences was not related to the number of cutpoints. Validation power closely approximated that predicted for a simple two group comparison. It is therefore recommended that exploratory analyses of prognostic factors formally employ some method of adjusting for increased Type I errors, such as independent validation sets, ad hoc adjustment factors, or other statistical methods of estimating the true risk.

Introduction

The literature is filled with newly identified prognostic factors. Many of these appear promising in the initial reports and then in follow-up studies fail to retain the same utility. Two examples from the breast cancer literature will illustrate the problem. First, there has been considerable interest and controversy regarding the relationship between the timing of surgery during the menstrual cycle of pre-menopausal breast cancer patients and their post-surgical prognosis. In about a dozen published reports, only about 25% found a significant relationship between cycle time and outcome [1]. Interestingly, some of the significant findings disagreed markedly on the optimal window for surgery. And second, in a recent review of the literature

Address for offprints: S. Hilsenbeck, UTHSCSA Med/Oncol, 7703 Floyd Curl Drive, San Antonio TX 78284-7884, USA.
[1] We regret to report that Dr. McGuire died on March 25, 1992, while this work was in progress.

on the prognostic value of HER-2/neu [2], studies involving mixed stages of breast cancer were about equally divided between negative ($p>0.05$) and positive ($p<0.05$) findings. Seven of eight studies of node-positive patients reported a significant effect for both relapse-free survival (RFS) and overall survival (OS), while in node-negative patients 1 of 11 studies found a highly significant effect for RFS and 4 of 11 for OS, respectively.

There are many possible explanations for why one study may fail to reproduce the results of another, among them technical differences in methodology (i.e. differences in antibodies used, or measurement of protein expression versus gene amplification); sampling heterogeneity or selection bias; or over-estimation of the true effect of the marker by the first study. This last explanation has not received much attention, and we will focus on it in this report.

Background

Prognostic factors are clinical or laboratory measurements that help predict clinical outcome. Typically, when a new factor is introduced, the utility of the factor is evaluated by retrospective analysis of survival or disease-free survival data from a group of patients on whom the factor has been measured. The investigators hypothesize that patients belong to two (or more) prognostically distinct groups, and that group membership is unknown, but related to the value of the prognostic factor of interest. Prognostic factor measurements usually take on a range of 2 or more possible values. The number of possible values can be small, as in the case of ploidy (diploid versus aneuploid), or very large, as with the measurement of estrogen receptor in fmol/mg protein. Factors measured by immunohistochemistry are often scored on an integer scale, e.g. ranging from 0 to 8. For simplicity, the investigators may hypothesize a dichotomous threshold effect for the prognostic factor, in which values below some cutpoint are associated with one prognosis and values above the cutpoint indicate a different prognosis. Statistical analysis is used to describe the underlying relationship between the numerical value(s) of the factor and outcome. Although some analysts have used the median value of the prognostic factor as the cutpoint, there is no *a priori* reason to expect that exactly half of the sample belongs to the good prognosis group and half belongs to the poor prognosis group. Instead, cutpoint analysis is used to obtain a maximum likelihood estimate of the cutpoint

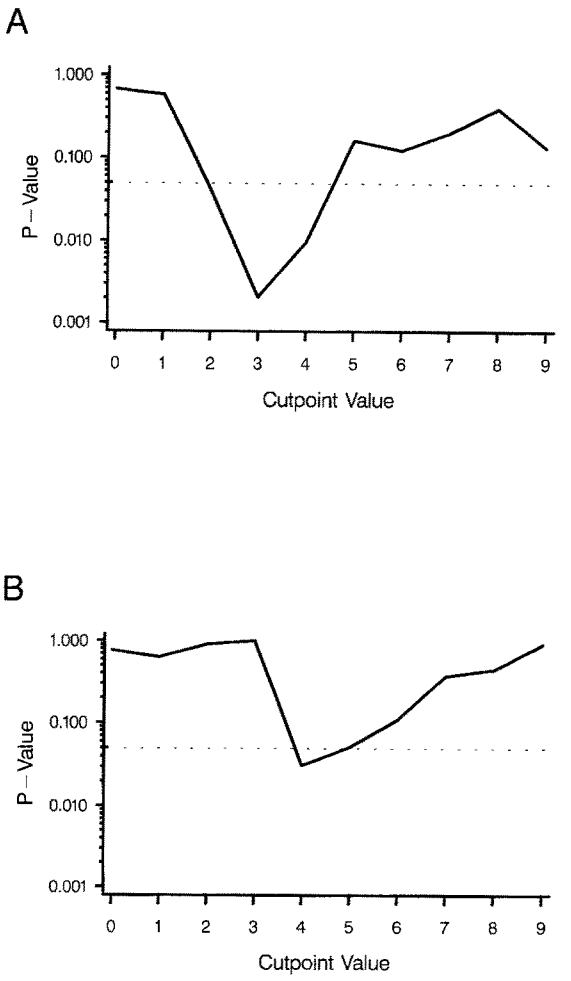

Figure 1. Cutpoint analysis curves for typical simulated datasets (n=250) with A) a true 10% difference in 5 year RFS, or B) no difference in 5 year RFS.

that best separates the groups. In this type of analysis, each prognostic value represented in the study sample is tried, in turn, as a cutpoint and the one that best separates the survival curves of the corresponding groups is selected. In recursive partitioning (see Albain et al. in this issue), the same approach is used to select successive splits. Several related algorithms have been used to perform the cutpoint selection. Log rank statistics, or equivalently the associated p-values, are often used as a measure of the separation of survival curves (Figure 1). A somewhat more efficient algorithm for cutpoint selection is based on the use of Martingale residuals from the null Cox Proportional Hazards regression model [3]. A modification to the "all possible Log rank tests" algorithm (above) was proposed by Abel, Berger, and Wiebelt [4], and is supposed to avoid bias that might be caused by uneven sample sizes. Cox regression has also been used to select optimal cutpoints, adjusted for other potentially important prognostic factors [5].

All of these methods are based on multiple looks at the data, and p-values associated with the Log rank statistics for the final selected cutpoint could be misleading. Consider the following scenario. A new prognostic factor takes on values from 0 to 9, and can therefore be dichotomized in 9 different ways (i.e. 0 vs 1-9; 0-1 vs 2-9; etc.). A group of 250 fictitious breast cancer patients is constructed such that patients with low values of the factor have about a 70% five-year relapse-free survival (RFS), while patients with high values have an 80% five year RFS. Based on a cutpoint analysis (Figure 1a), the best cutpoint is 3, with cases with low values having the better prognosis, and survival curves (Figure 2a) for the two groups (0-3 vs 4-9) exhibit a highly significant difference (p=0.001). Now consider a similar group of fictitious patients constructed so that the new factor is unrelated to RFS. Cutpoint analysis (Figure 1b) of this group selects 4 as the best cutpoint, and survival analysis (Figure 2b) finds a modest but apparently significant relationship. By random chance, high values are associated with good prognosis. However, using the selected cutpoint on an independent validation set of patients reveals the truth (Figure 2c) — there is no relationship between

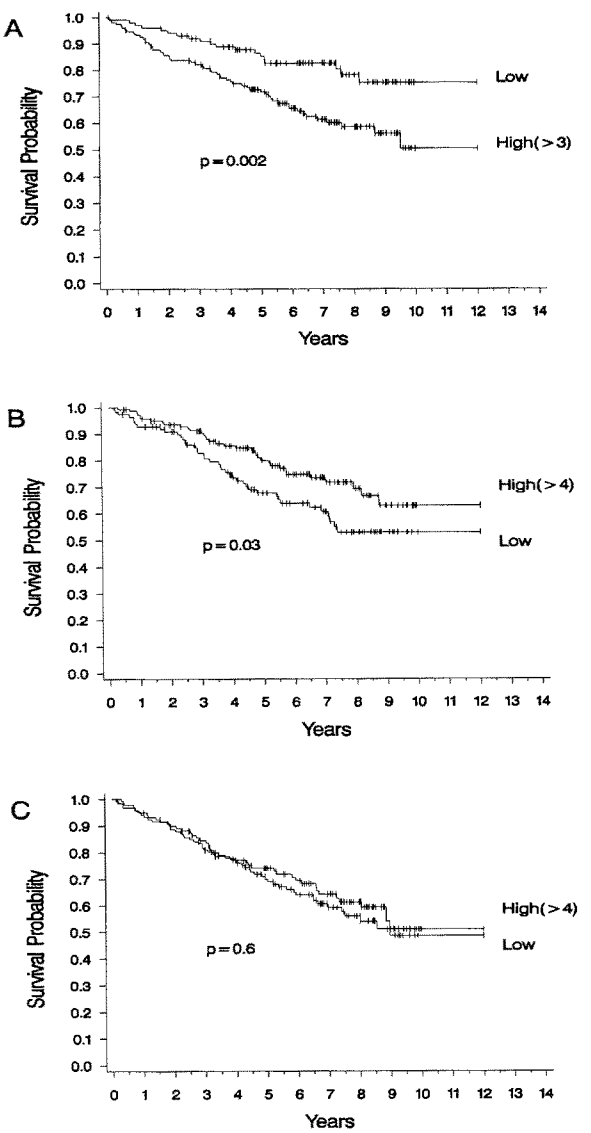

Figure 2. Relapse-free survival curves for simulated datasets (n=250). A) Training dataset with a true 10% difference in 5 year RFS; B) training dataset with no difference in 5 year RFS; and C) validation dataset associated with B.

the factor and RFS. Apparently we were misled by examining the data in too many different ways.

In concluding that there was a difference in prognosis when there was none, we have made a Type I error. Normally, we can control the risk of making such a mistake by setting the level of significance at which we make the test. We can report the p-value of the test (probability of obtaining a sample as unlikely or more unlikely than ours by chance alone), and conclude that the new factor is prognostically useful when the p-value falls below some pre-specified level (i.e. 0.05). When examining the same set of data in several different ways, the true overall p-value may be quite different from the p-value reported from single tests. This is sometimes called the "multiple comparisons problem".

Of course, we can also make a Type II error by failing to detect true differences in prognosis. The risk of Type II errors is normally controlled in clinical trials by determining the magnitude of a clinically useful difference and selecting a sample size that will detect that difference, if it exists, with reasonable certainty (i.e. power=0.80). Similar computations can be made for prognostic factor studies. Recently, we have recommended that prognostic factor studies be classified as pilot, definitive, or confirmatory, and that information from pilot studies be used to plan the size of definitive and confirmatory studies [6]. From a practical point of view, a study, particularly a pilot study, can be too powerful. Although not an error, detecting very small true differences with high sensitivity may not be desirable. Ideally, we would like to detect "important" differences perfectly and not detect "unimportant" differences at all.

The purpose of this paper is to investigate the problem of false-positive prognostic factor studies by examining the impact of several parameters, including number of cutpoints, sample size, and magnitude of true difference, on the outcome of cutpoint analyses.

Methods

A series of simulation experiments were undertaken to examine the effect of two factors, the number of cutpoints tested and the magnitude of the true difference in prognosis, on the outcome of cutpoint analysis. The number of cutpoints ranged from 1 to 50 (1,2,5,10,15,20,30,50), chosen to simulate the full range seen with real factors (i.e. ploidy, immunohistochemistry scores, S phase fraction, receptor levels, etc). The simulation of recurrence times was designed to approximate the real experience of node-negative breast cancer patients, with about a 70% five year RFS. Differences in prognosis were injected into the simulation by improving the five year RFS of the good prognosis group, up to 85% (70% [null], 71%, 73%, 76%, 80%, 85%). The examples described above represent single runs with 10 cutpoints and either 80% (10% difference) or 70% (no difference) five year RFS in the good prognosis group. For each individual experimental scenario, a set of fictitious patients was generated (500 or 250, to explore the effect of sample size) and then equally but randomly divided between a training set, which was used to select an optimum cutpoint using cutpoint analysis, and a validation set, which was used once to test the selected cutpoint. Each experimental scenario was run between 200 and 300 times. The overall experiment was evaluated to find the percentage of the runs for each combination of experimental factors that yielded significant (5% level) training and validation cutpoints, respectively.

Generation of recurrence time and prognostic factor data warrant further discussion. In most real investigations of prognostic factors it is not possible to follow all patients until failure. At the time of analysis, some patients have relapsed (or died, in the case of overall survival) and some patients are still disease free. The observed RFS times (and the associated disease status) can be thought of as resulting from two independent random processes, one governing time to recurrence

and the other governing time under observation (censoring time). When the time under observation is longer than the time to recurrence, the patient is counted as a recurrence at the recurrence time. When the time under observation is shorter than the recurrence time, the patient is censored. Normally, these two competing random processes are only hypothetical constructs that cannot be observed directly, but they are useful in generating simulated data. Of course, it is entirely possible to investigate the null (no difference in prognosis) case using real recurrence/censoring time data and randomly generated factor values. In fact, this approach was used in our investigation of menstrual cycle effects [1]. However, here we also wished to investigate the impact of cutpoint analysis on cases with varying magnitudes of true differences in prognosis. In order to do this it was necessary to generate time data that follow known distributions.

In our experiments, we used the Weibull distribution to generate recurrence times:

$$f(x;b,c) = \frac{c}{b}\left(\frac{x}{b}\right)^{c-1}\exp\left(-\left(\frac{x}{b}\right)^{c}\right)$$

In this case, the Weibull distribution has two parameters, a shape parameter c and a location parameter b. For c=1, the distribution is actually an exponential with a constant hazard rate of b. In this study, c was fixed at 1.1, which causes the hazard rate to increase slowly over time. Expected 5 year survival rates were adjusted by varying b from 12.75 (equivalent to a five year RFS of 70%) to 26.0 (equivalent to an 85% five year RFS). Values of the prognostic factor were generated first, so as to take on integer values ranging from 0 to the desired number of cutpoints. For a dichotomous factor like ploidy with a single cutpoint, the factor would take on values of 0 or 1. Fictitious patients with factor values below the midpoint were assigned b = 12.75, and patients with values above the midpoint were assigned a value for b corresponding to the required experimental scenario. The b's were then used to generate recurrence times for each patient. In this set of experiments, we did not study the effect of varying the location of the true cutpoint, nor did we examine the effect of several thresholds.

Finally, censoring times were generated independently for each fictitious patient as the maximum of two uniformly distributed random variables that could each range from 0 to 10 years. This generated a censoring distribution with relatively few early censorings, and more and more censoring at later times. This approximated our empirical experience and seemed to fit what might be expected in the clinic, where follow-up is most intense near the time of diagnosis and primary treatment and may tend to taper off as time goes by. Fictitious patients with recurrence times later than their censoring times were deemed censored at that point and all patients with recurrence times later than 10 years were censored at ten years. As might be expected in a comparable study of real node-negative breast cancer patients, about 70% of simulated cases were censored with a median follow-up of about 6 years. S Plus [7] was used to generate the simulated data and to compute the necessary statistics (i.e. Log rank tests, Kaplan-Meier curves).

Results

The results of the subset of runs for which the prognostic factor truly had no effect are summarized in Figure 3. As the number of tested cutpoints increased, so did the likelihood of obtaining a statistically significant result. For 10 cutpoints, the chances of making a Type I error and rejecting the null hypothesis at the nominal level of 5% was actually about 25%. That is, the true risk of a Type I error was about 5 times the nominal level. Doubling the sample size increased the risk slightly. With 50 cutpoints, the risk was about 38%, and in fact, based on another

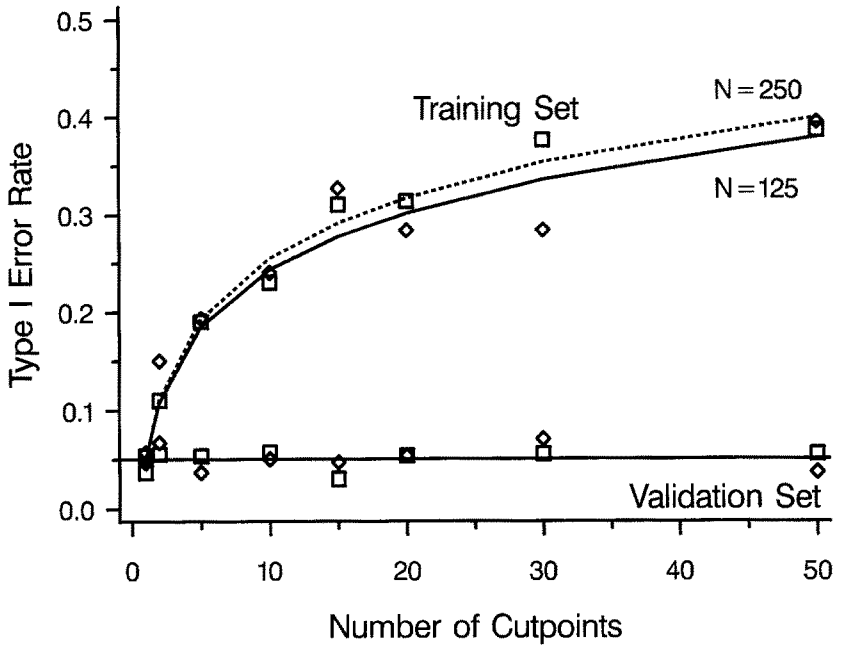

Figure 3. Type I error rates for training and validation datasets with no difference in prognosis. Observed rates for samples of 250 (□) and 125 (◇), and the corresponding fitted nonlinear regression lines (--- and —, respectively).

series of simulations not reported here, as the number of cutpoints increases toward the sample size (the upper limit, in the case of finite samples) the risk increases asymptotically to 50%. Note, however, that no matter how many cutpoints were tested in the training set, the proportion of cutpoints declared significant at the 5% level in the independent validation set was as steady as a rock at 5%. This is as expected.

As the prognostic factor took on true prognostic value, and the corresponding survival curves became increasingly separated, the results of the simulations produced a response or power surface (Figures 4-7). Figures 4 and 5 present the results

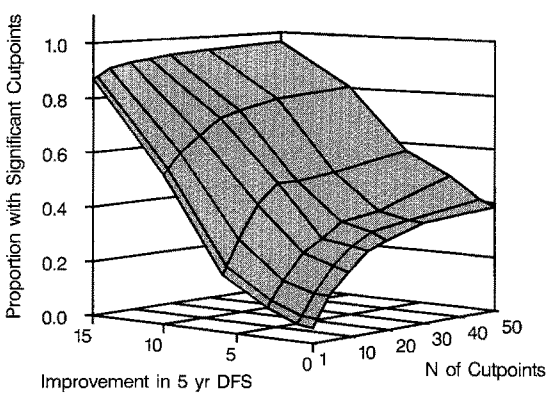

Figure 4. Observed power surface for training samples of 250 cases.

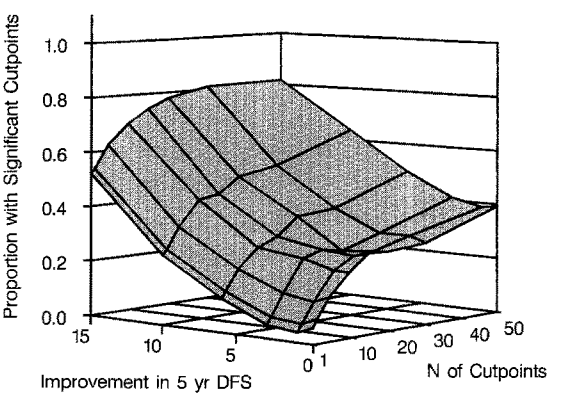

Figure 5. Observed power surface for training samples of 125 cases.

of cutpoint analyses on training datasets with 250 and 125 cases, respectively, while Figures 6 and 7 present the corresponding results for validation datasets. The rightmost edge of Figure 4 reproduces the dotted curve from Figure 5. Similarly, the rightmost edge of Figure 6 (and Figure 7) reproduces the validation line in Figure 3. Note that, in the training set, in addition to sample size, power was a function of both the magnitude of the true difference and the number of cutpoints tested. For a fixed improvement in 5 yr RFS, power increased with more cutpoints. When there was only a 6% difference in 5 year RFS (70% vs. 76%), but the factor had 50 possible cutpoints, we detected a statistically significant cutpoint 60% of the time. In contrast, power was less than 20% in the validation set. As expected, in the validation sets only sample size and the magnitude of the true difference had an effect on power. Indeed, the power to detect differences in the validation datasets was very close to that predicted using the method of George and Desu [8], an approximation that is often used to plan the size of clinical trials (Figure 8). The training power surfaces were always above (sometimes substantially above) the corresponding validation power surfaces, suggesting that with multiple cutpoints to choose from, the power to detect clinically insignificant differences may be substantial. It would not be surprising, therefore, if an exploratory study, testing several cutpoints, reported a very encouraging p-value while an independent study of similar size failed to obtain a significant p-value.

Discussion

It is clear from the numerical results that trying out even a small number of possible cutpoints, such as the median and quartiles of the putative prognostic factor, can significantly increase the true risk of a Type I error above the selected p-value. How can we protect ourselves? Several strategies seem reasonable.

First, and most obvious from this study, the use of an independent validation set provides almost certain protection. On testing the single selected cutpoint in a validation set, the risk of a Type I error was exactly what it should have been. Unfortunately, this is a potentially expensive, and in some cases an infeasible, form of insurance. It may be possible to improve things slightly by adjusting the relative sample sizes of the training and validation sets. A 50:50 allocation is probably not optimum. In fact, it may be preferable to use fewer cases in the training set and more cases in the validation set, in order to increase the power of the final validation test of the hypothesis. The precision of the estimate of the cutpoint, which is a function of training set size, may be less important than the power to

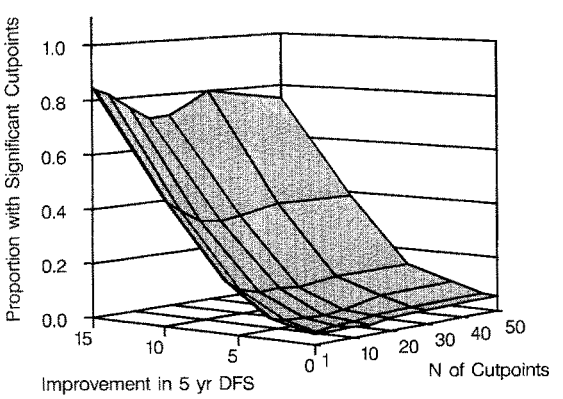

Figure 6. Observed power surface for validation samples of 250 cases.

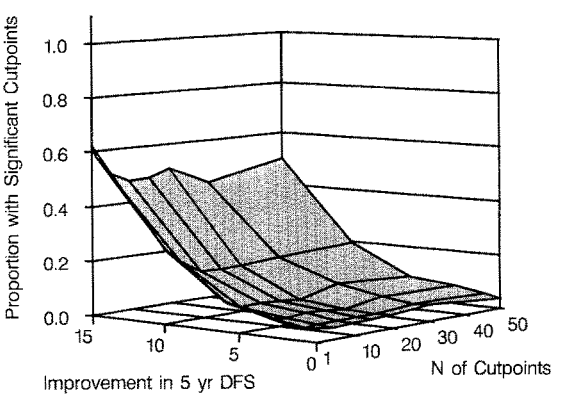

Figure 7. Observed power surface for validation samples of 125 cases.

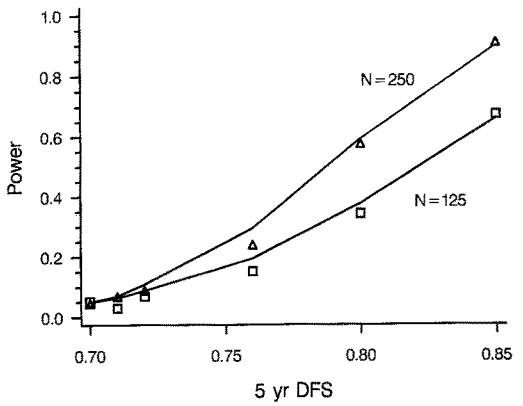

Figure 8. Predicted power for samples of 250 cases (△) and 125 cases (□).

detect useful factors after a reasonable cutpoint has been chosen.

Second, perhaps the level of significance at which each cutpoint is tested can be adjusted to account for the increase in risk due to testing multiple cutpoints, thus producing a more realistic description of the strength of the relationship between the new factor and prognosis. The 5% level of significance was chosen here for convenience. It is reasonable to expect that the magnification of the Type I error rate and the increase in power with increasing numbers of cutpoints would have been similar, proportionately, regardless of the selected level of significance. When using the 5% level as the critical value with 50 cutpoints, the true Type I error rate was about 40%. If the level were set at 1%, then we might expect to see apparently significant results (at that level) about 8% of the time. In fact, for sample sizes near those tested here, a good adjustment for the multiple cutpoint effect would be to use the fitted curves in Figure 3 to select a reduced "per cutpoint" level of significance, so that the true overall error rate would not exceed some desired level. For example, suppose a new factor was to be evaluated on 250 patients using an immunohistochemistry scoring system with 8 possible values, and therefore 7 possible cutpoints. The observed Type I error rate for this situation (from Figure 3) is about 21%. If we want the overall risk to be at most 5%, we could accomplish this by testing each of the 7 possible cutpoints at the 1.2% level (simulation adjusted level of significance = desired overall level * adjustment factor = 5%·[5%/21%]). Here, the adjustment factor is the significance level used to produce Figure 3 (5%) divided by the associated observed error rate (α_{obs} = 21%). In general, using the simulation data, the p-value required to declare any single cutpoint to be significant ($\alpha_{Simulation}$) will be:

$$\alpha_{Simulation} = \alpha \cdot \frac{0.05}{\alpha_{obs}}$$

where α is the desired overall level of significance.

Other methods of adjustment could also be used. The Bonferroni adjustment is computed by dividing the desired level of significance (0.05) by the number of comparisons that will be made:

$$\alpha_{Bonferroni} = \frac{\alpha}{k}$$

where $\alpha_{Bonferroni}$ is the p-value required to declare any single cutpoint to be significant, α is the desired overall level of significance, and k is the number of cutpoints to be tested. For example, if there are 10 possible cutpoints, testing each one at the 0.05/10 = 0.005 level will guarantee that the risk of erroneously declaring any cutpoint useful does not exceed 0.05. On the other hand, if it is reasonable to think of each cutpoint as an independent Bernoulli trial (which it is clearly not, since the same dataset is being reused and the results are therefore correlated), then the risk of a Type I error can be controlled by computing an independence adjusted level of significance:

$$\alpha_{Independent} = 1 - (1 - \alpha)^{\frac{1}{k}}$$

where $\alpha_{Independent}$ is the p-value required to declare any single cutpoint to be significant, α is the desired overall level of significance, and k is the number of cutpoints to be tested.

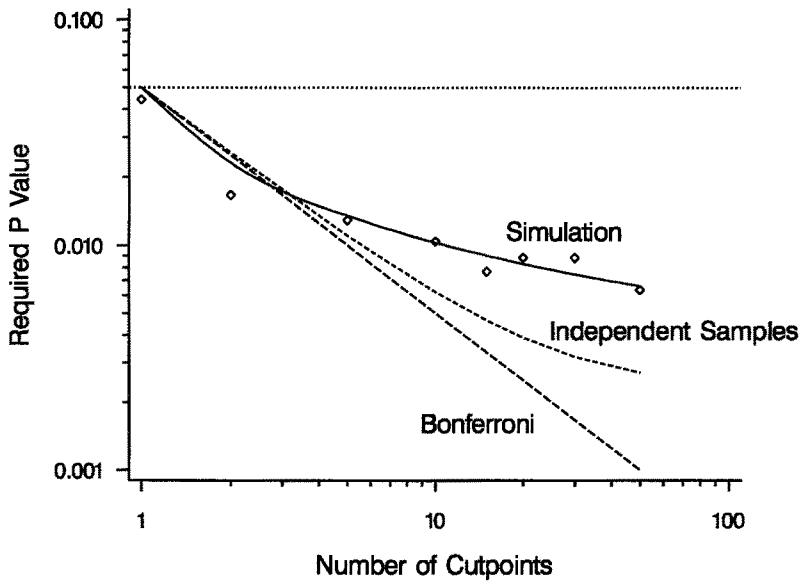

Figure 9. P values required to assure that the overall Type I error rate does not exceed 0.05, using three different methods. Simulation curve shows both the observed values (◇) and the fitted curve (—).

Although adjustment is often used to compensate for increased Type I error rates due to multiple testing, in cutpoint analyses the two traditionally used methods of adjustment both yield overly conservative values. Figure 9 illustrates the levels of significance that would be required by each method of adjustment in order to insure that the risk of making a Type I error will not exceed 5% for the entire analysis. All three methods produce similar critical values for 5 or fewer comparisons, but diverge markedly for higher numbers of cutpoints. Interestingly, the observed rate of Type I errors in a cutpoint analysis is much less than would be expected if independent samples were used. Apparently, the correlation between outcomes, from repeatedly testing the same data, works in our favor.

Finally, there are several more complex statistical approaches that may be useful in providing a more reliable picture of the true utility of a prognostic factor cutpoint. These include jackknife, bootstrap, and permutation test procedures. Although computationally intensive, permutation and resampling methods could provide a more data efficient method of screening prognostic factors. As suggested previously, definitive studies and confirmatory studies will no doubt still be required.

References

1. McGuire WL, Hilsenbeck SG, Clark GM: Optimal mastectomy timing. J Natl Cancer Inst 84:346-348, 1992.
2. Allred DC, Tandon AK, Clark GM, McGuire WL: HER-2/neu oncogene amplification and expression in human mammary carcinoma. *In* Pretlow TG II, Pretlow TP (eds) Biochemical and Molecular Aspects of Selected Cancers, Vol 1. Academic Press, 1991, pp 75-97.
3. Therneau TM, Grambsch PM, Fleming TR: Martingale residuals for survival models. Biometrika 77:147-160, 1990.
4. Abel U, Berger J, Wiebelt H: CRITLEVEL: An exploratory procedure for the evaluation of quantitative prognostic factors. Meth Inform Med 23:154-156, 1984.
5. Sigurdsson H, Baldetorp B, Borg Å, Dalberg M, Fernö, Killander D, Olsson H, Ranstam J: Flow cytometry in

primary breast cancer: improving the prognostic value of the fraction of cells in the S-phase by optimal categorization of cut-off levels. Brit J Cancer 62: 786-790, 1990.

6. McGuire WL: Breast cancer prognostic factors: Evaluation guidelines. J Natl Cancer Inst 83:154-155, 1991.
7. StatSci: S-PLUS Reference Manual, version 3.0. Statistical Sciences, Inc., Seattle WA, 1991.
8. George SL, Desu MM: Planning the size and duration of a clinical trial studying the time to some critical event. J Chron Dis 27: 15-24, 1974.