STATISTICS IN MEDICINE
Statist. Med. **17**, 1909–1922 (1998)

REGRESSION WITH BOUNDED OUTCOME SCORE: EVALUATION OF POWER BY BOOTSTRAP AND SIMULATION IN A CHRONIC MYELOGENOUS LEUKAEMIA CLINICAL TRIAL

A. TSODIKOV[1,2]* D. HASENCLEVER[2] AND M. LOEFFLER[2]

[1] *Huntsman Cancer Institute, University of Utah, 546 Chipeta Way, Salt Lake City, Utah 84108, U.S.A.*
[2] *Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Liebigstrasse 27, 04103 Leipzig, Germany*

SUMMARY

Evaluation of the treatment effect on cytogenetic ordered categorical response is considered in patients treated for chronic myelogenous leukaemia (CML) in a clinical trial initiated by the East German Group for Hematology and Oncology. A simulation model for the cytogenetic response (per cent of Philadelphia chromosome positive metaphases) serially measured in CML patients was constructed to describe roughly the sparse information available in medical literature. The model was used to construct a summary measure of response and to formulate the treatment effect as a regression with U-shape distributed ordered categorical data. Two simple models (vertical shift model and pooled conditional response model) were specifically designed to model the treatment effect 'observed' in a simulated 'pilot' data set. The powers were contrasted with the traditional proportional odds and binary models. The comparison was based both on repeated sampling from the simulated model and on bootstrap of 'given' pilot data set. We show that the specific models that address the treatment effect directly (as anticipated from pilot data) can gain in power as compared to the traditional proportional odds model when evaluated by bootstrap. However, the proportional odds model appears to be better with repeated sampling from the simulation model. To explain this discrepancy we generated 'pilot data sets' repeatedly from the simulation model and showed that the ordering of the bootstrap power estimates is unstable with reasonably complex models dependent on the random fall of the pilot data sets. This phenomenon clearly limits the usefulness of subtle modelling the form of the treatment difference observed in a small pilot data set. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

In planning a clinical trial a decision maker faces the problem of estimating the sample size and choosing the test which is most powerful in detecting the treatment effect. The decision is made under uncertainty, and at best a small pilot sample is available. One approach is to base the decision on bootstrapping the pilot data.[1,2] In planning a recent clinical trial we made an attempt to assess the reliability of this strategy by invoking the sparse information available in the medical literature to construct a reasonable simulation model of the trial.

* Correspondence to: A. Tsodikov, Huntsman Cancer Institute, University of Utah, 546 Chipeta Way, Salt Lake City, Utah 84108, U.S.A.

Contract/grant sponsor: Deutsche Forschungsgemeinschaft
Contract/grant number: Lo 342/6-1

CCC 0277–6715/98/171909–14$17.50

Received September 1996
Accepted October 1997

We consider a clinical trial comparing two treatment groups with bounded ordered categorical response U. The Wilcoxon two-sample test is the most popular test when the distributions (F and G) of the outcome score in the two treatment groups are not normal. Assuming a parametric representation of the set of alternatives it is possible to estimate the power of the test given the anticipated magnitude of the therapy effect in terms of the parameters. It is conventional to associate the Wilcoxon test with the location-shift Hodge–Lehmann[3] alternative

$$G(x) = F(x - \Delta). \tag{1}$$

However this is not applicable when the response is bounded.

Another suitable set of alternatives is given by the proportional odds model (McCullagh[4]) for ordered categorical response $U \in \{C_0, C_1, \ldots, C_k\}$, where $\{C_i\}_{i=0}^{k}$ are the possible categories of response, with C_i being better than C_j if $i < j$. Let F_i be the probability for a patient in the first treatment group to be in C_i or better, and let $\bar{F}_i = 1 - F_i$ be the probability to be in C_{i+1} or worse. Denote by G_i, $\bar{G}_i$ the similar probabilities in the second group, respectively. The proportional odds model specifies a parametric transformation of the form

$$\log(\bar{F}_i/F_i) = \log(\bar{G}_i/G_i) + \theta \tag{2}$$

where θ is the log odds ratio.

We will consider the categories based on a real valued score, so that the functions F and G are thought of as distribution functions (DF) of non-negative random variables U. The inference on θ can be based on the marginal likelihood ℓ_{marg}, the baseline function (F or G) being treated as nuisance. Let $x_1, \ldots, x_M; y_1, \ldots, y_N$ be the samples of U in the two groups, respectively. Let A_j be the set of ranks occupied by observations from group j in the joint order statistics $\xi_1, \ldots, \xi_{M+N}$, $(j = 1, 2)$. The marginal likelihood is defined as the probability to observe a given ordering of the data $\ell_{\text{marg}}(\theta) = \Pr\{A_1, A_2\}$ and it would be given by integration of the full likelihood over the subset of R^{M+N} on which $0 < \xi_1 < \cdots < \xi_{(M+N)}$, if the ordering were complete. Since we actually have ordinal (tied) data, only a partial ordering is available. The likelihood is then given by a sum of the terms corresponding to the likelihood under a complete ordering over all such complete orderings that are consistent with the observed partial ordering. The Wilcoxon test can be viewed as the score test for $H_0: \theta = 0$ with the semi-parametric proportional odds regression (2) based on the likelihood ℓ_{marg} (Jones and Whitehead[5]). We will follow this interpretation below.

In planning the chronic myelogenous leukaemia (CML) clinical trial (Section 2) we found that the anticipated distribution of response is U-shaped and bounded on [0, 1], the most important categories being $C_0: U = 0$ (complete responders) and $C_k: U = 1$ (non-responders). The problem of planning clinical trials with such response profile has been recently studied by Lesaffre *et al.*[1] and by Hilton.[6] It was reported that approximate formulae for the power based on the model (1) Lehmann[3] and (2) Whitehead[7] overestimate the power of the Wilcoxon test under the U-shaped response. A slight overestimation with the model (2) by the Whitehead method based on the score test was supposedly explained by the presence of scale effect in addition to the location shift. However, we have found (Section 5) that the ML estimates for the DF F and G based on the marginal likelihood fit the saturated curves perfectly, so that the proportional odds model captures the moderate scale effect in the U-shaped distribution of response, at least in this particular case.

Having constructed a simulation model of the CML trial (Section 3) we found that the therapy affects predominantly the extreme categories C_0 and C_k. Part of the reason for this effect is that an increase in therapy efficacy moves the patients to the left on the response axis while the response is

bounded. Consequently, the intermediate categories are subject to small changes as they experience both inflow and outflow, in contrast to the extreme categories. It seemed unnatural to express the effect as a combination of shift and scale transformations and we tried to address it directly (Section 4). The models are contrasted in Section 5.

2. MEDICAL BACKGROUND

CML is a clonal disorder of the primitive haemopoietic stem cell characterized by the presence of a chromosomal marker, the Philadelphia chromosome (Ph), in the leukaemic cells. In the past, the prognosis of patients with CML was poor. Until 1980 the focus of CML therapy was on haematologic remission (normalization of blood parameters) which was successful in 70 per cent to 80 per cent of patients. However, these remissions were mostly only symptomatic because cytogenetic studies in treated patients showed persistence of Ph-positive cells in most (> 90 per cent) of the marrow metaphases indicating the presence of residual disease in most patients (Kantarjian *et al.*[8]).

Recent advances in CML therapy are due to Interferon (IFN) that allowed for a cytogenetic response (< 90 per cent Ph+) in 40 per cent to 60 per cent of patients after prolonged application. The lowest percentage of Ph+ cells achieved by therapy has proved to be the strongest indicator for the prognosis. According to the protocol of the CML clinical trial initiated by the East German Group for Hematology and Oncology, those patients who fail to achieve complete (CR = 0 per cent Ph+) or partial (PR = 1 to 35 per cent Ph+) cytogenetic response during the first year of the therapy are classified as bad responders and who qualify for an intensified treatment.

All patients should have received intensive chemotherapy induction initially when entering the trial, followed by stem cell harvest from the peripheral blood. This manoeuvre aims to collect stem cells with less aggressive characteristics to be used later if the disease becomes more aggressive. The procedure is followed by the one-year therapy with cytostatics + IFN to maintain and improve the cytogenetic response. Typically, application of IFN decreases the percentage of Ph+ continuously. Bad responders after one year are treated by high dose chemotherapy with subsequent autologous stem cell transplantation in the hope of restoring the transient chronic phase of the disease.

The main focus of the trial is to contrast two modes of maintenance chemotherapy: IFN + HU (Hydroxyurea) versus IFN + ARA-C.

The main problem in evaluating the response lies in the fact that the true per cent of Ph+ cells remains latent. The estimation of the per cent of Ph+ cells by bone marrow biopsy and cytogenetic analysis is a costly and unpleasant procedure which can hardly be made more frequently than once every 4–6 months. At each evaluation some small number of metaphases k (usually $5 \leqslant k \leqslant 50$; 25 on average (Grossman *et al.*[9])) are extracted and the proportion v of the Ph-positive ones is determined. It is clear that v is a binomial frequency, the true proportion being its expectation. The necessity to make an early decision on the transplantation hinders a long-term follow-up of the marker.

Another complication is the loss of therapy efficacy due to tumour resistance to the drug or a discontinuation of the therapy due to toxicity. Both events provide an unobserved nadir in the trajectory of $v(t)$.

The evaluation of $v(t)$ is planned at $t = 0, 0{\cdot}5, 1$ (time in years). Bad responders ($v(0)$, $v(0{\cdot}5)$, $v(1) > 0{\cdot}35$) receive high dose chemotherapy with stem cell support. A similar strategy has been

used at the M.D. Anderson Cancer Center (Houston, Texas; Kantarjian *et al.*[10]). Evaluation of the efficiency of maintenance therapy is based on the summary measure

$$U_i = \min\{v(0),\ v(0{\cdot}5),\ v(1)\}, \quad i = \begin{cases} 1, & \text{IFN + HU} \\ 2, & \text{IFN + ARA - C} \end{cases} \tag{3}$$

with the distribution functions F and G corresponding to $i = 1, 2$, respectively. U assumes its values in one of the $k + 1$ categories $\{C_0, C_1, \ldots, C_k\} = \{0, 1/k, 2/k, \ldots, 1\}$. According to the protocol of the trial at least 25 metaphases should be analysed at each evaluation. It should be noted that an increase in k would result in a better expressed treatment effect with the same response pattern. Anticipating a slightly conservative inference, we assume $k = 25$ in the computations below.

3. SIMULATION MODEL

Based on observed trajectories $v(t)$ of the proportion of Ph+ cells in 8 CML patients published in Grossman *et al.*[9] and Kantarjian *et al.*[10] we decided to describe the data by a two-phase linear regression, truncated by the natural borders $v = 0$ and $v = 1$:

$$v(t) \sim \text{Bin}(k, \mu(t))$$

where

$$\mu(t) = \begin{cases} 0, & \mu^*(t) < 0 \\ \mu^*(t), & 0 \leqslant \mu^*(t) \leqslant 1 \\ 1, & \mu^*(t) \geqslant 1, \end{cases} \tag{4}$$

$$\mu^*(t) = \begin{cases} \mu(0) + (\lambda - \beta)t, & t \leqslant T \\ \mu(0) + (\lambda - \beta)T + \lambda(t - T), & t > T. \end{cases}$$

The parameter λ characterizes the rate of tumour growth, while β is a measure of efficacy of the IFN + HU/ARA-C therapy continuously applied in CML patients. The random variable T denotes the nadir point at which resistance starts to dominate. The slope of the trajectory before T is the superposition of the therapy effect and the tumour growth. After the therapy loses its efficacy (at time T) the slope is given by pure tumour growth. β, λ, T are thought of as gamma distributed random variables producing a random effects model. Some possible trajectories $\mu^*(t)$ are given in Figure 1.

The distribution of the random variable $\mu(0)$ was borrowed from the preliminary pilot data on patients, who have already passed the induction chemotherapy phase (Table I). The distributions of the other variables were fitted empirically to the information available in medical literature as given in Table II.

The variance of the random effects distributions was specified to fit the time to first cytogenetic response curve from Ozer *et al.*[11] That finally led to exponentially distributed random effects, which is perhaps due to high heterogeneity of the data. The mean value of the time to nadir was taken as 1·5 years, as estimated by a medical expert. Clinicians felt that the trial should have power to reveal an improvement in the 'true' success rate (probability of PR or CR) of 0·2 (an improvement from 0·35 to 0·54 is assumed). We consequently deduce the improvement in β from 0·45 to 1·125 to fit the above figures.

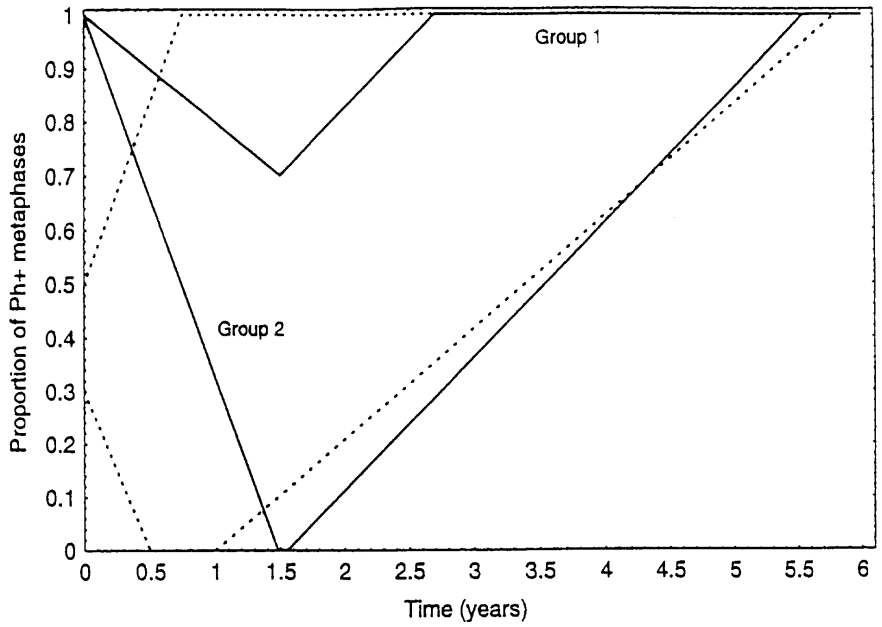

Figure 1. Possible trajectories of $\mu^*(t)$. Solid lines correspond to random effects variables set at their means: group 1 baseline therapy; group 2 improved therapy

Table I. The distribution of $\mu(0)$

Value of $\mu(0)$	0	0·17	0·82	0·97	1
Probability	0·03	0·13	0·22	0·19	0·44

Table II. Mean values of random effect parameters and probability of CR or PR as available from medical literature and the ones assumed in the model

Parameter	Simulation model	Kantarjian *et al.*[8]	Ozer *et al.*[11]	Kantarjian *et al.*[10]	Grossman *et al.*[9]
$\Pr(U \leqslant 0{\cdot}35)$	0·35	0·3–0·4	0·4	–	–
$\lambda - \beta$	− 0·2	–	–	− 0·2	− 0·2
λ	0·25	–	–	0·14	0·98

It is expected that the discontinuation of the therapy due to toxicity happens in 10 per cent of the patients of the IFN + HU group and in 30 per cent of the patients of the IFN + ARA-C and that toxicity is primarily attributed to HU or ARA-C. The time to such discontinuation is taken to be uniformly distributed in the interval [0, 6 months]. The risk of the discontinuation is considered as independent and competing with the time to that nadir. Patients with severe toxicities are supposed to change treatment arm. However, their responses are still attributed to their initial arms according to the intention to treat principle. Adding the above effects and taking account of the variance of U due to the evaluation procedure results in lower success rates: 0·27 (worse therapy) versus 0·42 (better therapy).

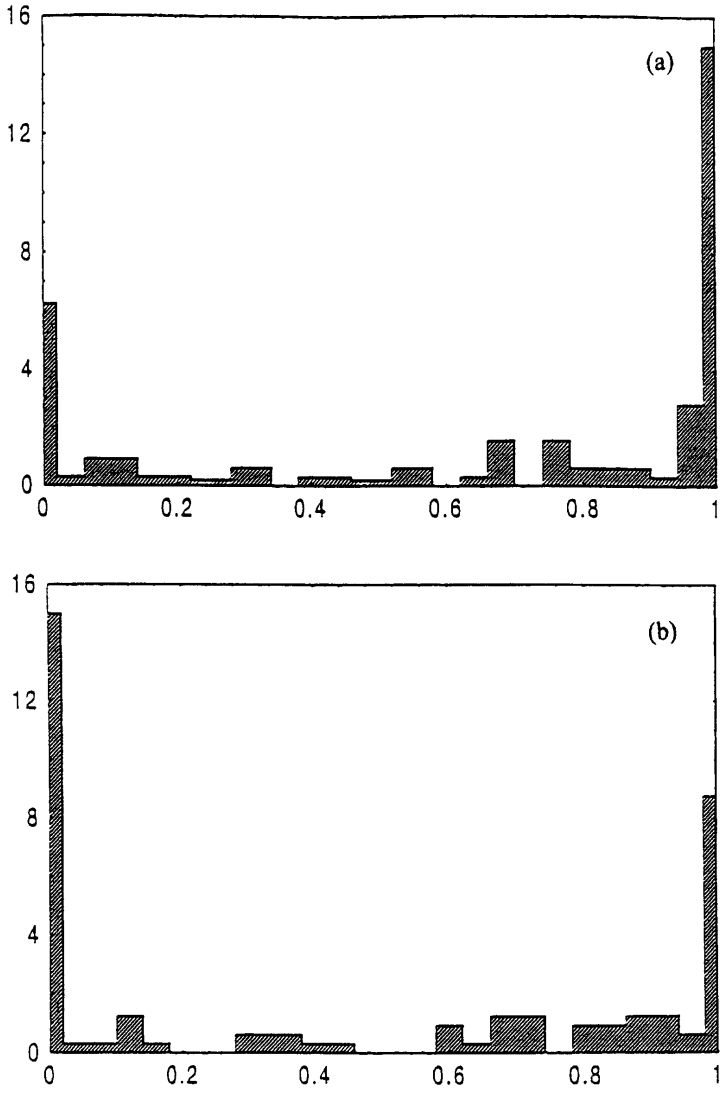

Figure 2. The response densities as generated by the simulation model: (*a*) baseline therapy; (*b*) improved therapy

The profile of the response distribution (F, G) is to a large extent influenced by the distribution of the proportion $v(0)$ of Ph+ cells in patients at time $t = 0$. Typical histograms of the response densities generated by the above model are given in Figure 2. They correspond to a sample of size 80 (each group) from the simulation model, which we later use as quasi pilot data to test the methodology. The 'true' distributions from the simulation model based on 5×10^5 replicates are shown in Figure 3. It is remarkable that the response profile is very similar to the one encountered by Lesaffre *et al.*[1] in another context. We note that the treatment effect is predominantly expressed in the increase in the number of CRs and a similar decrease in the number of non-responders, while the other categories are only slightly affected.

We considered it unreasonable to develop the testing procedure on the basis of the random effects model (4) although some related methods are feasible.[12,13] The model (4) is clearly overparameterized and we would have to make strong parametric assumptions to assure

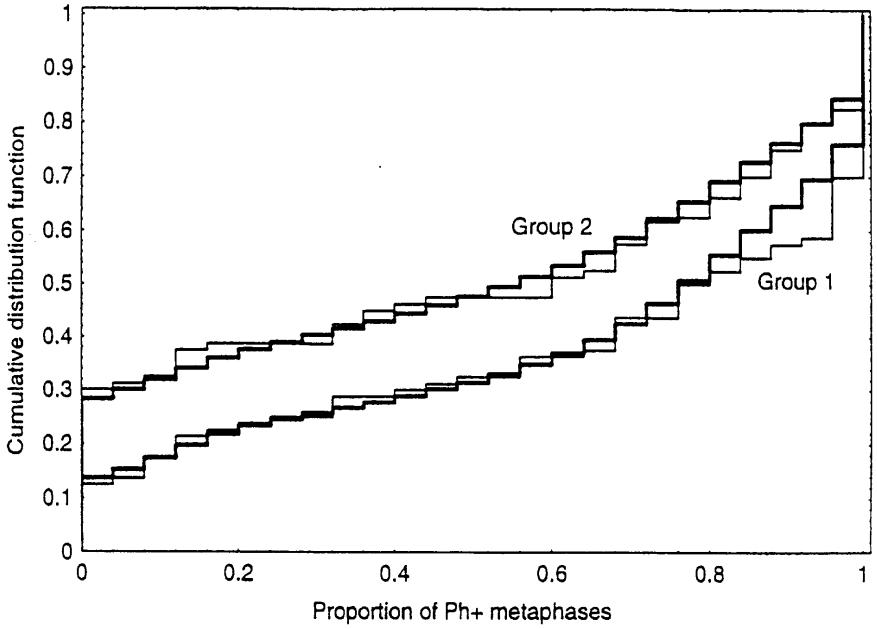

Figure 3. The 'true' distribution functions of the response as estimated by 5×10^5 simulations (thick lines) and the ones corresponding to quasi pilot data (a sample of size 80 (each group)) as given by the simulation model: group 1 baseline therapy; group 2 improved therapy

a reliable estimation procedure with just two highly correlated serial measurements per patient. Investing so much pre-knowledge seemed unwise, and we decided to use the summary measure U as suggested by Matthews *et al.*[14] and to apply 'surface' models to characterize the treatment effect on the summary measure.

4. REGRESSION MODELS

Given the distributions F and G and the numbers of patients m_i and n_i in each category C_i, in the two treatment groups, respectively, $i = 0, \ldots, k$, we can write the multinomial likelihood of the observed response as

$$\ell = \sum_{i=0}^{k} [m_i \log(\Delta F_i) + n_i \log(\Delta G_i)] \tag{5}$$

where $\Delta R_i = \Pr\{U \in C_i\}$ for $R = F, G$.

The saturated model is obtained by treating both functions F and G non-parametrically resulting in the likelihood ℓ_s and the obvious estimates $\Delta F_i = m_i/M$, $\Delta G_i = n_i/N$, $N = \sum_{i=0}^{k} n_i$, $M = \sum_{i=0}^{k} m_i$.

Under the homogeneity hypothesis H_0 we have the likelihood ℓ_p and the pooled estimates $\Delta F_i = \Delta G_i = (m_i + n_i)/(M + N)$.

4.1. Proportional odds model

A semi-parametric regression model specifies a parametric transformation $F \to G$, the baseline function F being treated non-parametrically. The proportional odds model (POM) (2) is the most

popular one with ordered categorical data. We rewrite the proportional odds model (2) in the form

$$\bar{G} = \frac{\bar{F}}{\eta_0 + \bar{F}(1-\eta_0)}, \ \eta_0 = \exp(\theta). \tag{6}$$

If θ is small and the sample is large, a score approximation can be used to obtain the estimate of θ in the form

$$\hat{\theta} = Z/V \tag{7}$$

where Z is the efficient score and V is the Fisher's information based on some likelihood of θ. Using the marginal likelihood ℓ_{marg} to avoid joint estimation of θ and F (or G) results in the following expressions:[7]

$$Z = \left.\frac{\partial \ell_{\text{marg}}}{\partial \theta}\right|_{\theta=0} = \frac{1}{M+N+1}\sum_{i=0}^{k} m_i(N_{0,i-1} - N_{i+1,k})$$

$$N_{a,b} = \sum_{i=a}^{b} n_i, \ N_{a,b} = 0, \ a > b$$

$$V = -\left.\frac{\partial^2 \ell_{\text{marg}}}{\partial \theta^2}\right|_{\theta=0} \approx \frac{MN(M+N)}{3(M+N+1)^2}\left\{1 - \sum_{i=1}^{k}\left[\frac{m_i+n_i}{M+N}\right]^3\right\}.$$

Alternatively we can fit the full model exactly by maximizing the likelihood (5). Usually, the software for logistic modelling based on the Newton–Raphson algorithm is used to fit the proportional odds model (for example, the SAS procedure LOGISTIC). We have used a simpler algorithm, reducing the problem to solving two nested algebraic equations. Proceeding from the likelihood equations:

$$\frac{\partial \ell}{\partial \eta_0} = \frac{N - n_k}{\eta_0} - \sum_{i=0}^{k-1}\frac{(n_{i+1}+n_i)F_i}{\bar{F}_i(1-\eta_0)+\eta_0} = 0 \tag{8}$$

$$\frac{\partial \ell}{\Delta F_j} = \frac{(m_j+n_j)}{\Delta F_j} - \frac{(m_k+n_k)}{\Delta F_k} - \sum_{i=j}^{k-1}\frac{(n_{i+1}+n_i)(1-\eta_0)}{\bar{F}_i(1-\eta_0)+\eta_0} = 0, \quad j = 0, \ldots, k-1 \tag{9}$$

define the function φ by:

$$\varphi(\Delta F_k) = \sum_{i=0}^{k} \Delta F_{k-i} \tag{10}$$

where ΔF_{k-i} are obtained recurrently from (9). It is easy to show by induction that φ is an increasing function. In addition we have $\varphi(0) = 0$, $\varphi(1) > 1$, so that the solution of the equation

$$\varphi(\Delta F_k) = 1 \tag{11}$$

exists and is unique. Proceeding from (11) we have the ΔF_i as functions of η_0 which reduces the problem to solving the algebraic equation (8) with respect to η_0. We can also substitute the approximate θ given by (7) in (9)–(11) to get the score approximation for the baseline function F.

4.2. Vertical shift model

Suppose that the therapy effect is associated with a flow of patients through the categories so that a fixed proportion η_v of patients in each category except C_0 are moved to a better category. This results in the distribution function G simply shifted vertically against F (vertical shift model, denote by VSM)

$$G(t) = F(t) - \eta_v, \quad 0 < t < 1. \tag{12}$$

The simple effect (12) is the first one which comes to mind when observing the distribution functions in Figure 3 and the similar functions in Lesaffre *et al.*[1] According to (12) the changes occur exclusively in the probabilities associated with extreme categories C_0 and C_k. Fitting the model (12) we get the likelihood ℓ_{VSM} and the estimates

$$\begin{aligned} \Delta F_i &= (m_i + n_i)/(M + N), \quad i = 1, \ldots, k-1 \\ \Delta F_0 &= \frac{m_0}{m_0 + m_k} \frac{m_0 + n_0 + m_k + n_k}{M + N} \\ \Delta F_k &= \frac{m_k}{m_0 + m_k} \frac{m_0 + n_0 + m_k + n_k}{M + N}. \end{aligned} \tag{13}$$

The derivation of these estimates as well as those of (15) (see below) from the score equations is straightforward but cumbersome and therefore not presented here.

The likelihood ratio statistics $D = 2(\ell_{\text{VSM}} - \ell_p)$ of the homogeneity hypothesis is expressed as a function of the numbers of patients in the extreme categories because the estimates ΔF_i, $i = 1, \ldots, k-1$ coincide with that of the saturated model under the homogeneity hypothesis.

4.3. Pooled conditional response model

The VSM (12) can be relaxed by assuming independent changes in the probabilities associated with C_0 and C_k and the identity of the conditional distributions of the response in the two treatment groups, given the intermediate response $(C_1, \ldots, C_{k-1})$. We have the pooled conditional response model (denoted by PCRM)

$$\begin{aligned} \Delta G_0 &= \Delta F_0 + \Delta_0; \ \Delta G_k = \Delta F_k - \Delta_k \\ \frac{\Delta F_i}{1 - \Delta F_0 - \Delta F_k} &= \frac{\Delta G_i}{1 - \Delta G_0 - \Delta G_k}, \quad i = 1, \ldots, k-1. \end{aligned} \tag{14}$$

Fitting the model (14) results in the likelihood ℓ_{PCRM} and the estimates

$$\Delta F_i = \frac{m_i + n_i}{M\left(1 + \dfrac{N_{1,k-1}}{M_{1,k-1}}\right)}, \quad \Delta G_i = \frac{m_i + n_i}{N\left(1 + \dfrac{M_{1,k-1}}{N_{1,k-1}}\right)}, \quad i = 1, \ldots, k-1, \tag{15}$$

$$N_{1,k-1} = \sum_{i=1}^{k-1} n_i, \quad M_{1,k-1} = \sum_{i=1}^{k-1} m_i$$

while the estimates $\Delta F_0, \Delta F_k, \Delta G_0, \Delta G_k$ coincide with that of the saturated model. Again, the likelihood ratio statistic $D = 2(\ell_{\text{PCRM}} - \ell_p)$ is related to the extreme categories.

Also, the simple binary approach and the Pearson's χ^2 statistic[15] will be used to test the homogeneity of the two groups with the response U_i, $i = 1, 2$ dichotomized into the two categories ($U_i \leqslant 0{\cdot}35$) versus ($U_i > 0{\cdot}35$).

It should be noted that other effects might be considered by multiplying the baseline density by some specified function as suggested by Lesaffre *et al.*[1] or by introducing a shift and scale transformation of the baseline distribution as suggested by Hilton.[6] In fact the possibilities to refine the model and improve its fit on the basis of the available pilot data are without limit. In the next section we show that this strategy is unreasonable since the reproducibility of a test's superiority is low. We contrast the models POM (6), VSM (12), PCRM (14) and the binary model using resampling from the Figure 2 as well as repeated sampling from the simulation model (Section 3) directly.

5. NUMERICAL EXPERIMENTS

Since the patients are entering the subsequent trial which focuses on the effect of stem cell transplantation, the estimate of the baseline function is considered as equally important to predict the number of patients to be transplanted after 1 year.

The regression models were fitted to the quasi pilot data (80 patients each group) corresponding to Figure 2 as shown in Figure 4. Seemingly, all the models capture the effect equally well. However, bootstrapping the data with 10^5 replicates shows superiority of the models (12) and (14) over the traditional proportional odds model (6) as given in the Table III. We have tuned the number of replicates to attain an accuracy of 2×10^{-3} in terms of the size of 95 per cent confidence interval for estimating power from bootstrapping the pilot data.

It should be noted that the score method by Whitehead[7] provides a conservative estimate for the exact power based on the empirical distributions from Figure 2 in our particular case. For example, for 80 patients we have the estimate 0·750 (compare with Table III).

We have repeated the analysis basing the sampling on the original simulation model. Although the realization shown in Figure 2 was typical, it was not fully representative of the set of alternatives provided by the simulation model (Figure 3). As shown in Table IV, the proportional odds model turns out to be superior with repeated sampling from the simulation model. Unsurprisingly, the binary approach is the worst in both settings.

We found that the ordering of the power estimates is unstable between repeated sampling from the simulation model and bootstrapping the pilot data. Repeated sampling (200 samples, 80 patients each) from the simulation model and bootstrapping each such 'pilot' sample yields the box plots of the power differences by models shown in Figure 5. We note that except for the differences involving the binary model all other plots are approximately symmetric about zero. This means the probability of the right choice of the model on the basis of pilot data is about 0·5 in our particular case, and we could just toss a coin.

In general the power of a statistical test depends on the fit of the underlying model to the 'true' curves. It is common to characterize this unobserved fit by the mean prediction error (estimated by cross-validation).[16] Asymptotically the leave one out cross-validation is equivalent to the Akaike's information criterion[17] AIC $= -2\ell + 2$ (numbers of parameters), so that with large samples and known number of parameters, cross-validation is not truly necessary. It is interesting whether the ordering of powers can be predicted in a similar way. With our quasi pilot data we found that the ordering of the POM (AIC = 869·89) versus VSM (AIC = 868·50) and of the PCRM (870·10) versus VSM could have been predicted while a reverse ordering of the POM

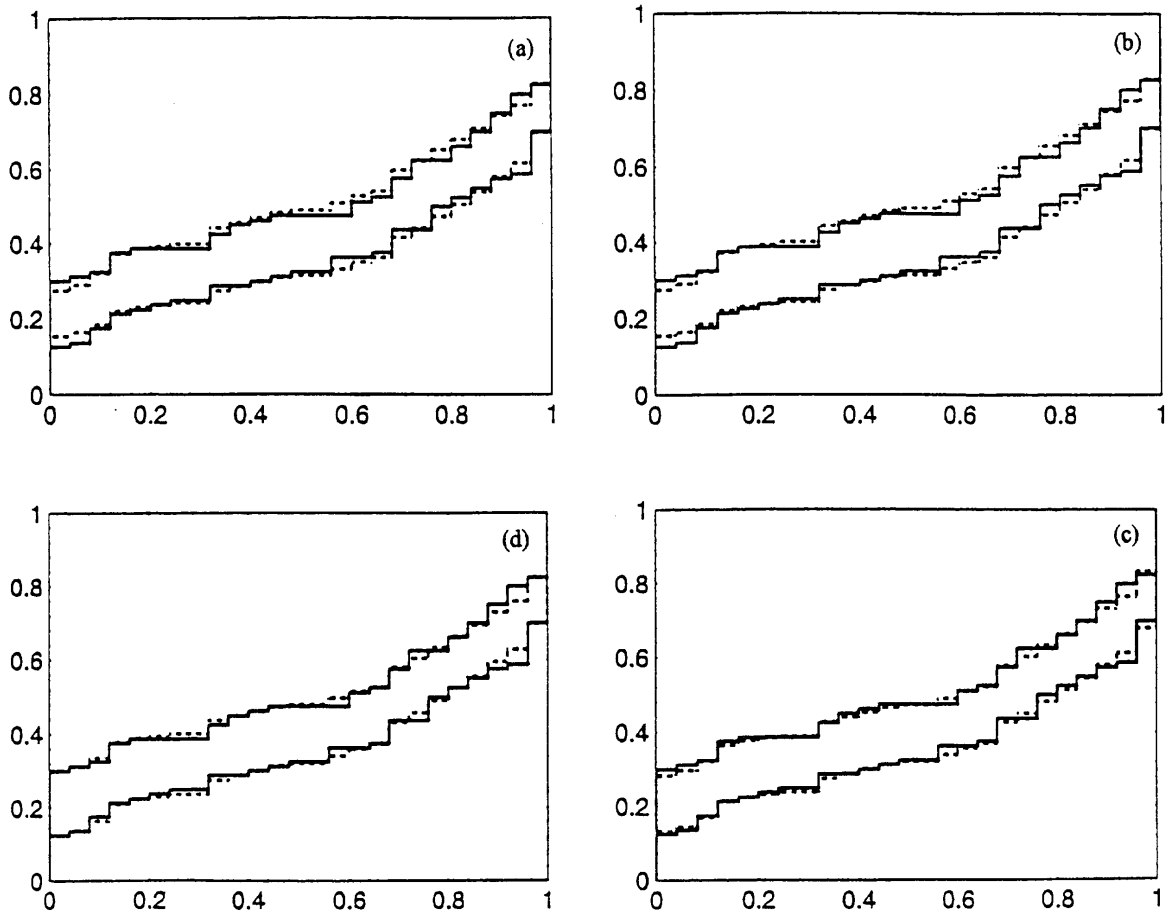

Figure 4. Fitted curves according to models POM (6), VSM (12) and PCRM (14). Solid line is the survivor function as specified by the density shown in Figure 2. Dotted lines relate to the fitted regression models: (*a*) score approximation, POM; (*b*) POM, jointly fitted baseline distribution and treatment effect; (*c*) vertical shift model (12); (*d*) pooled intermediate response model (14)

Table III. Bootstrap estimates of power of the likelihood ratio test based on the models POM, VSM, PCRM, the score test (Wilcoxon) based on POM and the Pearson's χ^2 test

Model	Number of patients		
	60	80	100
Proportional odds, score	0·639	0·761	0·846
Proportional odds, exact	0·639	0·761	0·846
Vertical shift	0·720	0·831	0·904
Pooled conditional response	0·640	0·772	0·860
Binary χ^2	0·351	0·452	0·529

versus PCRM was observed. However, the discrepancy between the power estimates in the latter case is too small to be interpreted. Even if an asymptotic criterion for the choice of a test becomes available, the bootstrap will still retain its importance because the size of pilot data is usually quite small.

Table IV. Estimates of power of the likelihood ratio test based on the models POM, VSM, PCRM, the score test (Wilcoxon) based on POM and the Pearson's χ^2 test. Repeated sampling from the simulation model of Section 3

Model	Number of patients		
	80	100	120
Proportional odds, score	0·663	0·758	0·830
Proportional odds, exact	0·666	0·760	0·831
Vertical shift	0·631	0·727	0·802
Pooled conditional response	0·579	0·682	0·765
Binary χ^2	0·526	0·611	0·691

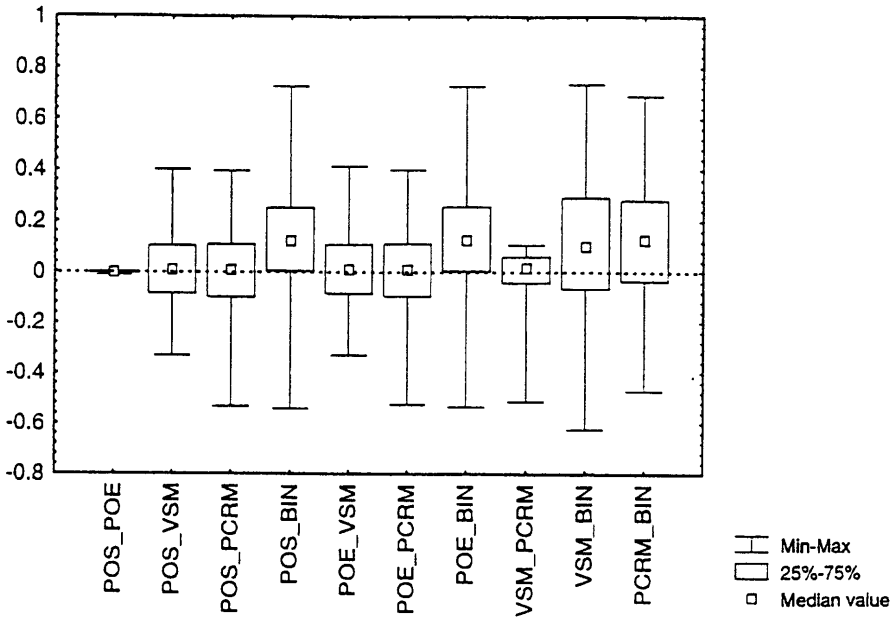

Figure 5. Box plots of 200 paired differences between the bootstrap power estimates obtained in two steps: 1. Generate pilot data (80 patients in each group) from the simulation model; 2. Estimate the powers based on the pilot sample and construct the paired differences with POS (proportional odds model, score method), POE (proportional odds model, exact solution of the likelihood equations), VSM, PCRM or binary models

6. DISCUSSION

In planning a recent clinical trial we tried to anticipate the type and size of the treatment effect and to choose a test to compare the two treatment groups. An easy decision would have been just to choose the most popular Wilcoxon test. However this seemed unwise because of the suspicion that the proportional odds model behind the test might not adequately capture the treatment effect, resulting in biased and underpowered inference. If a pilot data set were available one could try anticipating the outcome of the trial by bootstrapping based on the pilot data. In the absence of a reliable pilot data set, we constructed a simulation model to fit the information available in medical literature. The model was used to generate the outcome of a trial, which followed the

U-shaped bounded response distribution, the treatment predominantly affecting the extreme categories of the response.

We constructed the VSM and PCRM models to address the effect directly and compared them with the traditional proportional odds and binary models. Bootstrapping the 'observed' outcome, we found that the tests based on the suggested models outperformed the traditional ones. With VSM and PCRM it is easier to quantify the treatment effect by specifying the changes in the numbers of complete responders and of non-responders, while with the traditional models the whole distribution of the response has to be anticipated. However, the proportional odds model was found to be superior when we used the simulation model (and not the 'observed' realization) to replicate the experiment.

By generating 'pilot data sets' from the simulation model we showed that the ordering of the bootstrap power estimates can be unstable, dependent on the random fall of the pilot data sets, and that the estimates based on the models of reasonable complexity (POM, VSM and PCRM) are similar in this respect. This phenomenon clearly limits the usefulness of subtle modelling the form of the treatment difference observed in a small pilot data set.

Even if a pilot data set is available we would recommend simulation invoking additional information from the medical literature to assess the reproducibility of pilot data and the stability of power estimates based on it. At the same time caution has to be exercised in interpreting the simulation model because its adequacy is difficult to test from the data.

ACKNOWLEDGEMENTS

The authors are thankful to Professor Dr. W. Helbig, the supervisor of the CML clinical trial, and Dr. Krahl for stimulating discussions and placing the clinical data and relevant medical literature at our disposal. This research is supported by the grant Lo 342/6-1 of the Deutsche Forschungsgemeinschaft.

REFERENCES

1. Lesaffre, E., Scheys, I., Fröhlich, J. and Bluhmki, E. 'Calculation of power and sample size with bounded outcome scores', *Statistics in Medicine*, **12**, 1063–1078 (1993).
2. Hamilton, M. A. and Collings, B. J. 'Determining the appropriate sample size for nonparametric tests for location shift', *Technometrics*, **33** (3), 327–337 (1991).
3. Lehmann, E. L. *Nonparametrics, Statistical Methods Based on Ranks*, Holden Day, San Francisco, 1975.
4. McCullagh, P. 'Regression models for ordinal data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
5. Jones, D. R. and Whitehead, J. 'Sequential forms of the log rank and modified Wilcoxon tests for censored data', *Biometrika*, **66**, 105–113 (1979); Correction *Biometrika*, **68**, 576 (1981).
6. Hilton, J. 'The appropriateness of the Wilcoxon tests in ordinal data', *Statistics in Medicine*, **15**, 631–645 (1996).
7. Whitehead, J. 'Sample size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
8. Kantarjian, H., Deisseroth A., Kurzrock, R., Estrov, Z. and Talpaz, M. 'Chronic myelogenous leukemia: a concise update', *Blood*, **82**, 691–703 (1993).
9. Grossman, A., Silver, R., Szatrowski, T., Gutfriend, A., Verma, R. and Benn, P. 'Densitometric analysis of Southern Blot autoradiographs and its application of monitoring patients with chronic myelogenous leukemia', *Leukemia*, **5**, 540–547 (1991).
10. Kantarjian, H., Talpaz, M., Keating, M., Estey, E., O'Brien, S., Beran, M., McCredie, K., Gutterman, J. and Freireich, E. 'Intensive chemotherapie induction followed by interferon-alpha maintenance in patients with Philadelphia chromosome-positive chronic myelogenous leukemia', *Cancer*, 1201–1206 (1991).

11. Ozer, H., George, S., Schiffer, C., Rao, K., Rao, N., Wurster-Hill, D., Arthur, D., Powell, B., Gottlieb, A., Peterson, B., Rai, K., Testa, J., LeBeau, M., Tantravahi, R. and Bloomfield, C. 'Prolonged subcutaneous administration of recombinant 2b Interferon in patients with previously untreated Philadelphia chromosome-positive chronic-phase chronic myelogenous leukemia: effect on remission duration and survival: Cancer and Leukemia Group B Study 8583', *Blood*, **82**, 2975–2984 (1993).
12. Hinkley, D. 'Inference about the intersection in two-phase regression', *Biometrika*, **56**, 495–504 (1969).
13. Stephens, D. 'Bayesian retrospective multiple-changepoint identification', *Applied Statistics*, **43**, 159–178 (1994).
14. Matthews, J., Altman, D., Campbell, M. and Royston, P. 'Analysis of serial measurements in medical research', *British Medical Journal*, **300**, 230–235 (1990).
15. Everitt, B. S. *The Analysis of Contingency Tables*, 2nd edn, Chapman and Hall, 1992.
16. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, 1993.
17. Stone, M. 'An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion', *Journal of the Royal Statistical Society, Series B*, **39**, 44–47 (1997).