Biometrika (1977), **64**, 2, *pp.* 191–9
Printed in Great Britain

Group sequential methods in the design and analysis of clinical trials

BY STUART J. POCOCK
Medical Computing and Statistics Group,
Medical School, University of Edinburgh

SUMMARY

In clinical trials with sequential patient entry, fixed sample size designs are unjustified on ethical grounds and sequential designs are often impracticable. One solution is a group sequential design dividing patient entry into a number of equal-sized groups so that the decision to stop the trial or continue is based on repeated significance tests of the accumulated data after each group is evaluated. Exact results are obtained for a trial with two treatments and a normal response with known variance. The design problem of determining the required size and number of groups is also considered. Simulation shows that these normal results may be adapted to other types of response data. An example shows that group sequential designs can sometimes be statistically superior to standard sequential designs.

Some key words: Clinical trial; Group sequential method; Repeated significance test.

1. INTRODUCTION

In most randomized clinical trials, patient entry is sequential so that results become available sequentially. Both medical ethics and the natural curiosity of investigators require an ongoing assessment of the accumulating data to see if a treatment difference is sufficient to stop the trial. Formal sequential methods (Armitage, 1975) have been applied in only a small fraction of actual trials. This lack of use arises because few trials conform to the usual sequential design with two treatments, patient entry in matched pairs, instantaneous patient evaluation, a normal or binary response and continuous surveillance of accumulating data.

Thus, most clinical trials proceed with a rather ill-defined stopping rule. Periodic analyses of the accumulating data using standard significance tests may provide some objectivity though such repeated testing can greatly increase the null probability of detecting a significant difference at some point (Armitage, McPherson & Rowe, 1969).

However, Armitage (1975, p. 27) shows that repeated significance testing can be a useful sequential method. The idea is that after every observation, or matched pair of observations if there are two treatments, one carries out a two-sided significance test so that if a 'nominal' significance level α' is achieved one stops the trial declaring evidence of a treatment difference. Here α' and the maximum number of observations, or pairs, N are chosen so that the 'overall' significance level α, that is the probability of detecting a treatment difference under the null hypothesis, and the power $1-\beta$, that is the probability of detecting a treatment difference under some alternative hypothesis H_A, are set to some required levels. Thus, given α, H_A and β one can use numerical methods to determine α' and N, though the precise results depend on the type of response variable and significance test.

Such sequential designs tend to have the limitations mentioned earlier. In particular the need for continuous assessment after every observation can be especially difficult to organize so that a natural adaptation is to apply the significance test at longer equally-spaced intervals,

say every $2n$ patients in the case of two treatments. Let N be redefined as the maximum number of consecutive patient groups of size $2n$, the notation otherwise being unchanged. Then for any particular response variable and significance test one can determine numerically appropriate values of α', n and N; here α' is a function of α, n and N which leaves two quantities n and N to be determined from α, H_A and β. There is a consequent flexibility of design enabling one to make a convenient choice for either $2n$, the number of patients per group, or N, the maximum number of groups.

The group sequential designs for a normal response with known variance defined in § 2 can be adapted to provide a good approximation for other types of response as is shown in § 3. Further practical details are considered in § 4.

2. GROUP SEQUENTIAL DESIGNS FOR A NORMAL RESPONSE WITH KNOWN VARIANCE

Consider a clinical trial with two treatments A and B in which a homogeneous sample of patients is entered sequentially. Suppose treatment assignment is by a randomized permuted block design so that each consecutive group of $2n$ patients has n on each treatment. Let response be a normal random variable with known variance σ^2 and unknown means μ_A and μ_B for treatments A and B respectively. Let $\bar{x}_{Aj}$ and $\bar{x}_{Bj}$ be the observed mean responses on treatments A and B in the jth group of n patients each. Then,

$$\bar{d}_i = \sum_{j=1}^{i} (\bar{x}_{Aj} - \bar{x}_{Bj})/i$$

is normal with mean $\delta = \mu_A - \mu_B$ and variance $2\sigma^2/(in)$ so that a two-sided significance test of the null hypothesis $\mu_A = \mu_B$ applied after i groups has significance level

$$P_i = 2[1 - \Phi\{\sqrt{(in)}\,\bar{d}_i/(\sqrt{2}\sigma)\}].$$

Now, from § 1 a group sequential design requires that if $P_i < \alpha'$, some nominal significance level, one would stop the trial, claiming evidence of a treatment difference. One also requires a maximum number of groups N whereby one claims no evidence of a treatment difference if $P_i > \alpha'$ for all $i = 1, \ldots, N$. Here α' and N are chosen so that the overall significance level α for the null hypothesis $\mu_A = \mu_B$ has some desired value, say 0·05 or 0·01. In fact, given N and α the value of α' does not depend on n since $\{\bar{x}_{Aj} - \bar{x}_{Bj}\}$ $(j = 1, \ldots, N)$ are identically distributed normal mean differences with known variance $2\sigma^2/n$. Such repeated testing on accumulating normal data is described by Armitage *et al.* (1969). However, they dealt with quite high values of N suitable for sequential designs whereas in group sequential designs one is interested in smaller values of N. Hence, based on the same method of numerical quadrature as Armitage *et al.*, Table 1 shows values of α' and its corresponding standardized normal deviate for $\alpha = 0{\cdot}05$ or 0·01 and N in the range 2 to 20. It is interesting that for $\alpha = 0{\cdot}05$ and $N = 11$ the nominal significance level $\alpha' = 0{\cdot}01$.

In choosing appropriate values of n and N there may be external influences such as a predetermined cost or length of trial to fix nN or a natural interval between analyses to fix n. However, it is useful to evaluate the statistical power achieved by the possible designs under consideration.

Now, for identically distributed normal observations with unknown mean δ and known variance σ^2 the power $1 - \beta$ of detecting δ significantly different from 0, using any predefined test procedure, is a function of δ/σ. Group sequential testing of normal data is based on $\bar{x}_{Aj} - \bar{x}_{Bj}$ $(j = 1, \ldots, N)$ with means $\mu_A - \mu_B$ and variances $2\sigma^2/n$. Hence, the power of detecting a difference $\delta = \mu_A - \mu_B$ is a function of $(\sqrt{n}\delta)/(\sqrt{2}\sigma) = \Delta$, say. Given Δ, N, α and α', one can apply numerical quadrature to compute $1 - \beta$. In practice it is more useful to know the value

of Δ corresponding to particular β, N, α and α'. This inverse calculation uses an iterative linear interpolation convergence procedure applied to the numerical quadrature values of $\beta(\Delta, N, \alpha)$. Table 2 shows such values of for various α, N and $1-\beta$. Number of groups $N = 1$ is included to enable comparison with fixed sample designs.

Also, it is easy to compute the average number of groups before termination of trial when the alternative hypothesis $\mu_A - \mu_B = \delta$ is true; see Table 3.

Table 1. *The nominal significance level α' and its corresponding standardized normal deviate z for use in normal group sequential testing, with known variance, for various values of number of groups N and overall significance level α*

	$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}01$			$\alpha = 0{\cdot}05$		$\alpha = 0{\cdot}01$	
N	α'	z	α'	z	N	α'	z	α'	z
2	0·0294	2·178	0·0056	2·772	9	0·0112	2·535	0·0019	3·099
3	0·0221	2·289	0·0041	2·873	10	0·0106	2·555	0·0018	3·117
4	0·0182	2·361	0·0033	2·939	11	0·0101	2·572	0·0017	3·133
5	0·0158	2·413	0·0028	2·986	12	0·0097	2·585	0·0016	3·147
6	0·0142	2·453	0·0025	3·023	15	0·0086	2·626	0·0015	3·182
7	0·0130	2·485	0·0023	3·053	20	0·0075	2·672	0·0013	3·224
8	0·0120	2·512	0·0021	3·078					

Table 2. *Values of $\Delta = (\sqrt{n}\delta)/(\sqrt{2}\sigma)$ for normal group-sequential tests with various values of overall significance level, α, power, $1-\beta$, and maximum number of groups, N*

	$\alpha = 0{\cdot}05$					$\alpha = 0{\cdot}01$				
$(1-\beta)$	0·5	0·75	0·9	0·95	0·99	0·5	0·75	0·9	0·95	0·99
N										
1	1·960	2·634	3·242	3·605	4·286	2·576	3·250	3·858	4·221	4·920
2	1·477	1·967	2·404	2·664	3·152	1·921	2·405	2·839	3·099	3·584
3	1·243	1·647	2·007	2·221	2·622	1·607	2·006	2·362	2·575	2·973
4	1·096	1·449	1·763	1·949	2·297	1·413	1·760	2·070	2·255	2·600
5	0·994	1·311	1·592	1·759	2·071	1·277	1·588	1·866	2·032	2·341
6	0·916	1·207	1·464	1·617	1·903	1·175	1·460	1·714	1·866	2·149
7	0·855	1·125	1·364	1·506	1·770	1·095	1·359	1·595	1·735	1·998
8	0·805	1·058	1·282	1·415	1·662	1·030	1·277	1·498	1·630	1·875
9	0·764	1·002	1·214	1·339	1·573	0·975	1·209	1·417	1·541	1·773
10	0·728	0·955	1·156	1·275	1·497	0·929	1·150	1·348	1·466	1·686
11	0·697	0·914	1·105	1·219	1·431	0·888	1·100	1·289	1·401	1·611
12	0·670	0·878	1·061	1·170	1·373	0·853	1·056	1·237	1·344	1·502
15	0·605	0·791	0·956	1·053	1·235	0·769	0·950	1·112	1·209	1·389
20	0·529	0·691	0·835	0·919	1·077	0·671	0·829	0·970	1·053	1·209

The main use of Tables 2 and 3 is to obtain suitable values of n and N given α, δ and $1-\beta$. Note that for any chosen value of N, $n = 2\ (\Delta\sigma/\delta)^2$, so that it is convenient to express δ in terms of σ. Also, n will need to be rounded to the nearest integer resulting in a slight change in β which should be negligible in practice. Consider the following example.

Example. Let $\alpha = 0{\cdot}05$, $\delta = 0{\cdot}5\sigma$ and $1-\beta = 0{\cdot}90$. Then for $N = 5$ groups, $\Delta = 1{\cdot}592$ from Table 2, so that $n = 2 \times (1{\cdot}592/0{\cdot}5)^2 = 20{\cdot}27$. From Table 3, under the alternative hypothesis $\delta = 0{\cdot}5\sigma$ the average number of patients on the inferior treatment $= 2{\cdot}838 \times 20{\cdot}27 = 57{\cdot}5$. This is not exact since n must be rounded to an integer, but the error is slight. This example is repeated for various values of N in Table 4.

As N increases there is an increase in nN, the maximum number of each treatment, from 84 in a fixed design to about 111 in a 20 group design. However, the main objective of a group sequential procedure is to reduce the number of patients on an inferior treatment by early termination of trial. This effect is illustrated by the average number of patients per treatment under the alternative hypothesis $\delta = 0{\cdot}5\sigma$ falling from 84 for the fixed design to about 56 for a 20 group design. One achieves a substantial drop by adopting a 2 or 3 group design with little extra advantage gained in 10 or 20 group designs. Furthermore, the equivalent sequential design based on repeated significance tests, as described by McPherson & Armitage (1971),

Table 3. *The average number of groups to termination of trial under the hypothesis* $\mu_A - \mu_B = \sqrt{2}\Delta\sigma/\sqrt{n}$, *where* Δ *is obtained from Table 2 as a function of the maximum number of groups,* N, *overall significance level,* α, *and power,* $1-\beta$

	$\alpha = 0{\cdot}05$					$\alpha = 0{\cdot}01$				
$(1-\beta)$	0·5	0·75	0·9	0·95	0·99	0·5	0·75	0·9	0·95	0·99
N										
2	1·76	1·58	1·41	1·31	1·16	1·80	1·64	1·47	1·37	1·21
3	2·52	2·19	1·88	1·71	1·44	2·60	2·30	2·00	1·83	1·55
4	3·28	2·80	2·36	2·12	1·74	3·40	2·96	2·53	2·29	1·90
5	4·04	3·41	2·84	2·53	2·05	4·19	3·62	3·06	2·75	2·27
6	4·80	4·02	3·32	2·94	2·36	4·99	4·27	3·59	3·22	2·63
7	5·56	4·63	3·80	3·35	2·68	5·78	4·93	4·12	3·68	2·99
8	6·31	5·24	4·28	3·77	2·99	6·57	5·58	4·65	4·14	3·35
9	7·07	5·85	4·76	4·18	3·31	7·37	6·23	5·18	4·61	3·72
10	7·83	6·46	5·24	4·59	3·62	8·16	6·89	5·71	5·07	4·08
11	8·58	7·07	5·72	5·01	3·93	8·95	7·54	6·24	5·53	4·44
12	9·34	7·68	6·20	5·42	4·25	9·74	8·20	6·77	6·00	5·03
15	11·60	9·50	7·63	6·66	5·19	12·12	10·16	8·36	7·39	5·89
20	15·38	12·54	10·03	8·72	6·76	16·07	13·42	11·00	9·70	7·67

Table 4. *Maximum and average number of patients under* H_A *for several values of* N *in a normal group sequential design with* $\alpha = 0{\cdot}05$, $\delta = 0{\cdot}5\sigma$ *and* $1-\beta = 0{\cdot}90$

Number of groups, N	1	2	3	5	10	20
Required no. per group, n, on each treatment	84·1	46·2	32·2	20·3	10·7	5·6
Maximum no. of patients on each treatment, nN	84·1	92·4	96·6	101·5	106·9	111·4
Average no. of patients on each treatment under the alternative hypothesis	84·1	65·2	60·5	57·5	56·0	55·9

has a maximum of 121 patients on each treatment with an average number of 58·3 patients if $\delta = 0{\cdot}5\sigma$. This result for $n = 1$ and $N = 121$, was obtained using a computer program written by C. K. McPherson. This sequential design is inferior to group sequential designs with $N = 5$, 10 or 20 in that both the maximum and average numbers of patients are greater under both null and alternative hypotheses.

This example provides some general insight into the properties of normal group sequential testing. In general, a group sequential design with even a quite small number of groups provides a substantial reduction in average sample size when treatment differences exist. In fact, such a reduction may often be close to or even better than that achieved by standard sequential designs. Thus, if continual analysis is burdensome there is apparently little loss of sensitivity in applying a group sequential stopping rule instead. Also, there will be increased

robustness to departures from normality as a result of using group means and more time to prepare for and reflect upon one's analysis with consequent improvement in data quality and interpretation of results.

3. Group sequential designs for other types of response variable

For a normal response with known variance, α' and Δ depend only on N, α and β. However, for other response variables, e.g. normal with unknown variance, binomial or exponential, α' and Δ also depend on n, so that tabulation of α' and Δ would be a much greater task. However,

Table 5. *Point estimates of* α, $1-\beta$ *for group-sequential two-sample* t *test using values of* α', Δ *from normal group-sequential tests; based on* 5000 *simulations in each case*

N	n	Required α	Point estimate of α using α' from Table 1*	Required $1-\beta$	Point estimate of $1-\beta$ using α', Δ from Tables 1–2*
2	5	0·05	0·0478	0·5	0·447
				0·9	0·854
		0·01	0·0084	0·5	0·388
				0·9	0·801
2	20	0·05	0·0504	0·5	0·481
				0·9	0·890
		0·01	0·0106	0·5	0·466
				0·9	0·882
5	5	0·05	0·0474	0·5	0·460
				0·9	0·876
		0·01	0·0104	0·5	0·434
				0·9	0·855
5	20	0·05	0·0526	0·5	0·493
				0·9	0·888
		0·01	0·0104	0·5	0·485
				0·9	0·885
10	5	0·05	0·0476	0·5	0·476
				0·9	0·894
		0·01	0·0084	0·5	0·464
				0·9	0·881
10	20	0·05	0·0536	0·5	0·492
				0·9	0·888
		0·01	0·0122	0·5	0·483
				0·9	0·886

* For point estimates near 0·01, 0·05, 0·5 and 0·9 the 95 % confidence limits are approximately 0·003, 0·006, 0·014 and 0·008 respectively, either side of the point estimate.

we now show by simulation that the values of α' and Δ in Tables 1 and 2 can be adapted for other types of response.

(*a*) *Normal response with unknown variance.* Consider the problem in § 2 except with σ^2 unknown. Then, one stops the trial after the ith group if the significance level P_i of the two-sample t statistic on $2(in-1)$ degrees of freedom is less than some value α'. Given N and n, can the values of α' and Δ in Tables 1 and 2 provide good approximations to the required α and β? Table 5 shows point estimates for α and β, each based on the proportion of 5000

simulations with a significant group-sequential test. Using α' in Table 1 results in estimated α close to the required levels of 0·05 and 0·01. One expects some loss of power with variance unknown, but in practice $1-\beta$ is not seriously underestimated.

(*b*) *Exponential response.* For two samples of size k from the same exponential distribution, the ratio of sample means has an F distribution on $(2k, 2k)$ degrees of freedom, which leads to a standard F test for comparing two exponential means. Thus, for an exponential response we consider a group sequential F test with α' as in Table 1. For all 12 combinations of $N = 2$, 5 or 10, $n = 5$ or 20, and required $\alpha = 0{\cdot}05$ or 0·01, estimates of the true overall significance level, obtained from 5000 simulations, were within the range of chance variation of the required α.

If $\bar{y}_A$ and $\bar{y}_B$ are means from samples size k of exponential distributions with means λ_A and λ_B respectively, then $\log(\bar{y}_A/\bar{y}_B)$ is approximately $N\{\log(\lambda_A/\lambda_B), 2/k\}$. Thus, using the notation of § 2, a normal approximation to the exponential replaces $\delta = \mu_A - \mu_B$ by $\log(\lambda_A/\lambda_B)$ and σ by 1. For particular values of N, n, α and β one can determine approximate values of

$$\Delta = (\tfrac{1}{2}n)^{\frac{1}{2}} \log(\lambda_A/\lambda_B)$$

as defined in Table 2 which enables some assessment of power for a group sequential F test.

For example, if $N = 5$, $n = 20$, $\alpha = 0{\cdot}05$ and $1-\beta = 0{\cdot}5$ then, from Table 2, $\Delta = 0{\cdot}994$ so that $\lambda_A/\lambda_B = 1{\cdot}37$. From 5000 simulations the estimate of actual power was 0·503, near the intended value of 0·5. Other simulation examples showed similarly accurate power calculations.

(*c*) *Binary response.* Consider a binary variable, response versus nonresponse, and suppose that after k groups there are r_A and r_B responses out of nk patients on treatments A and B respectively. Then,

$$\sqrt{2}\,\frac{\{r_A(nk-r_B) - r_B(nk-r_A)\}}{\{nk(r_A+r_B)(2nk-r_A-r_B)\}^{\frac{1}{2}}} = U_k$$

has approximately the distribution $N(0, 1)$ under the null hypothesis of no treatment difference. Consider a group sequential design with U_k as test statistic and two-sided nominal significance levels α' as in Table 1. The accuracy of this normal approximation has been studied for all 48 combinations of $N = 2$, 5 or 10, $n = 5$, 10, 20 or 50, $\alpha = 0{\cdot}05$ or 0·01, and null hypothesis probability of response $\pi = 0{\cdot}5$ or 0·3. Estimates of the true overall significance level were from 5000 simulations.

For $n = 5$ patients per group on each treatment such a design is not too accurate; e.g. for required $\alpha = 0{\cdot}05$ estimates ranged from 0·040 for the case $N = 2$, $n = 5$, $\pi = 0{\cdot}3$ to 0·063 for $N = 10$, $n = 5$, $\pi = 0{\cdot}3$. Also, for $N = 2$ groups one needed $n = 50$ or more for an estimate close to the required α; for example $N = 2$, $n = 50$, $\pi = 0{\cdot}5$, with required $\alpha = 0{\cdot}05$, has an estimated true $\alpha = 0{\cdot}049$ which becomes 0·056 if $n = 20$ instead. However, $N = 5$ or 10 groups gave accurate results provided $n \geqslant 10$; for example $N = 5$, $n = 10$, $\pi = 0{\cdot}5$ had estimates 0·049 and 0·011 for required $\alpha = 0{\cdot}05$ and 0·01 respectively.

Neither inclusion of the Yates continuity correction nor adaptation of the Fisher–Irwin exact test are as accurate as the above group sequential test based on U_k; for example for $N = 5$, $n = 20$, $\pi = 0{\cdot}5$ and required $\alpha = 0{\cdot}05$ the estimated true α falls from 0·048 to 0·026 if the Yates correction is used. This is in accord with the known inappropriateness of the continuity correction when multiple test statistics are combined.

Now we examine the power of binary group-sequential testing based on U_k. Let response probabilities on A, B be π_A and π_B and let $\bar{\pi} = \frac{1}{2}(\pi_A + \pi_B)$. Then, from the usual normal approximation we replace $\Delta = \sqrt{n}(\mu_A - \mu_B)/(\sqrt{2}\sigma)$ in § 2 by $\Delta = \sqrt{n}\,(\pi_A - \pi_B)/\{2\bar{\pi}(1-\bar{\pi})\}^{\frac{1}{2}}$. Given n, N, α, β one can find Δ from Table 2 to get values of π_A and π_B. With prior knowledge of one treatment π_A could be fixed, leaving a quadratic equation for π_B.

For example, if $N = 5$, $n = 20$, $\alpha = 0{\cdot}05$, $\beta = 0{\cdot}5$ and $\pi_A = 0{\cdot}5$ we obtain $\Delta = 0{\cdot}994$ and $\pi_B = 0{\cdot}6552$ or $0{\cdot}3448$. From 5000 simulations the estimated true power is $0{\cdot}503$. Several other examples gave reliable estimates of β provided that the design also gave a good estimate of α.

(*d*) *The Wilcoxon test.* Simulation of an exact group sequential Wilcoxon test involves excessive computing so we consider the following normal approximation instead. Let R_A be the sum of ranks on treatment A after $k \leqslant N$ groups. Then one stops the trial if

$$\frac{|R_A - \frac{1}{2}nk(2nk+1)|}{kn(\frac{1}{6}nk+\frac{1}{12})^{\frac{1}{2}}}$$

exceeds the normal deviate z defined in Table 1. Simulation is still lengthy, but a few examples have been studied and are shown in Table 6. Results are reasonably good and one presumes similar success is achieved for designs with larger values of N and/or n.

Table 6. *Repeated Wilcoxon test. Estimated α based on* 5000 *simulations*

N	2	3	5	2	3	5
n	20	10	10	20	10	10
Required α	0·05	0·05	0·05	0·01	0·01	0·01
Estimated α	0·0512	0·0444	0·0482	0·0088	0·0084	0·0084

Using the group sequential Wilcoxon test on normal data can lead to only a slight loss in power. For example, for $N = 2$, $n = 20$ and $\alpha = 0{\cdot}05$ a difference in treatment means of $0{\cdot}467\sigma$ is detected with power $0{\cdot}5$ for a normal group sequential test. The Wilcoxon equivalent has estimated power $0{\cdot}452$.

The above examples justified by simulation lead one to suspect that normal group sequential methods could also be adapted to many other types of response variable. It would be useful if more theoretical results could be developed to justify the general validity of such normal theory for even quite small samples.

4. Further comments

4·1. *Variation in size of group*

It may be impracticable to have exactly equally-spaced analyses of the data, so that the kth analysis may not have exactly kn patients on each treatment. For instance, consider a trial with two treatments, with purely random assignment, a Poisson process rate λ for patient entry and group sequential analysis at equally spaced intervals of time T. Suppose that data are normal with unknown variance, the expected number of patients per treatment in each group being $\frac{1}{2}\lambda T = 20$ and there being $N = 5$ groups. For $\alpha = 0{\cdot}05$ the regular group sequential t test should have $\alpha' = 0{\cdot}0158$ from Table 1 and with power $1-\beta = 0{\cdot}5$ or $0{\cdot}9$ one should have $\Delta = 0{\cdot}994$ or $1{\cdot}592$ respectively from Table 2. For this more irregular design, estimates of α and β based on 500 simulations are $0{\cdot}0516$ and $0{\cdot}490$, $0{\cdot}888$ respectively. Similar satisfactory results may be expected for other slightly irregular patterns of group sequential analysis.

4·2. *Stratification*

When patient factors may influence response one needs to stratify both in design and analysis. Sometimes one anticipates some interaction between treatment and strata so that each stratum essentially forms a clinical trial in itself with a separate design and analysis.

Alternatively, and perhaps more commonly, one expects some variation in response between

strata but with a constant treatment difference. Consider an additive model for a normal response with unknown variance, two treatments and M strata; let the response y_{ij} on treatment i within stratum j be defined by $y_{ij} = t_i + s_j + e$, where $\{t_i\}$ $(i = 1, 2)$ and $\{s_j\}$ $(j = 1, \ldots, M)$ represent treatment and stratum effects and e is normal with mean 0 and unknown variance σ^2. The parameters are uniquely defined if we set $t_2 = -t_1$. A fixed sample test for a treatment difference, is described by Armitage (1971, pp. 264–8) and we now consider adapting this test to a group sequential design.

Consider a clinical trial with random patient entry, each patient having an equal chance of falling into each of M strata. Suppose treatment allocation is either (i) purely random or (ii) in random permuted blocks of two in each stratum. Suppose one has N groups of $2n$ patients such that after each group one stops the trial if the above test is significant at a two-sided nominal level α' determined from Table 1. For instance, if $M = 8$ strata, $N = 5$ groups and $2n = 40$ patients per group then $\alpha' = 0{\cdot}0158$. One hopes that the resultant overall significance level $\alpha = 0{\cdot}05$ and that the treatment difference $\delta = t_1 - t_2$ detected with power $1-\beta$ can be obtained from $\Delta = (\sqrt{n}\delta)/(\sqrt{2}\sigma)$ given in Table 2. These conjectures are confirmed by the estimates in Table 7 of actual α, β from 5000 simulations, using α', Δ from Tables 1 and 2.

Table 7. *Estimates of overall significance level α and power $1-\beta$ for purely randomized and stratified designs*

	Purely random assignment	Random permuted blocks of 2 in each stratum
Estimated α; required value, 0·05	0·0498	0·0542
Estimated $1-\beta$; required value, 0·5	0·466	0·499
Estimated $1-\beta$; required value, 0·9	0·880	0·897

Evidently, normal group sequential designs can be adapted successfully to cope with stratification though for purely random assignment the possibility of unequal treatment numbers within strata may account for a slight loss of power. Simulations with 2, 4 and 16 strata gave similar results.

One anticipates similar success with group sequential designs for (i) other types of response variable and stratification and (ii) allowance for quantitative patient variables using analysis of covariance.

4·3. *Other than two treatments*

With three or more treatments one treatment may be a standard control in which case one can define a separate group-sequential design for each other treatment compared with the control. Alternatively, if one wishes to consider all treatments on equal terms some global significance test, say an F ratio test for one-way analysis of variance if response is normal, could be adapted to group-sequential methods.

Group-sequential methods can also be used in an experiment with only one treatment in which the response results are compared with a known standard. For instance, for a normal response with known variance and null hypothesis mean μ_0 the values of α' and Δ in Tables 1 and 2 still apply except that Δ becomes $\sqrt{n}\delta/\sigma$, where δ is the difference between μ_0 and the mean under an alternative hypothesis.

4·4. *Other types of group-sequential design*

The methods described in this article are not based on any formal optimal properties such as minimizing sample size under a particular hypothesis, but they are a rational, easy-to-use

form of stopping rule. Canner (1976) paper, has suggested varying α' over the N tests, possibly having more stringent tests early on in a trial, but it is unclear what statistical advantages this might have. Also, one could use one-sided rather than two-sided tests, which involves some reformulation of Tables 1–3, but most clinical trials require two-sided alternatives.

Elfring & Schultz (1973) consider the properties of certain designs for a binary response and two treatments but their choice of design appears somewhat arbitrary.

In many trials there is some time lapse between patient entry and the observation of response. Difficulties arise in applying sequential methods to such follow-up studies (Armitage, 1975, Chapter 7), and further research is needed to see if group sequential methods can be adapted.

References

Armitage, P. (1971). *Statistical Methods in Medical Research.* Oxford: Blackwell.

Armitage, P. (1975). *Sequential Medical Trials.* Oxford: Blackwell.

Armitage, P., McPherson, C. K. & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A **132**, 235–44.

Canner, P. L. (1976). Repeated analysis of clinical trial data. In *Proc.* 9th *Int. Biometrics Conf.*, Vol. 1, pp. 261–75. Raleigh, N. Carolina: Biometric Soc.

Elfring, G. L. & Schultz, J. R. (1973). Group sequential designs for clinical trials. *Biometrics* **29**, 471–7.

McPherson, C. K. & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J. R. Statist. Soc.* A **134**, 15–25.

[*Received June* 1976. *Revised October* 1976]