CONTROLLED CLINICAL TRIALS IN CANCER RESEARCH

EVA SKOVLUND

Knowledge of important aspects of the design and analysis of clinical trials is essential to clinical researchers and readers of medical literature. A brief description of proper trial design, including the contents of a trial protocol, as well as different strategies to avoid bias, is given. The concept of p-values is explained, and some commonly used statistical analysis methods are mentioned. Statistical power is defined, and two useful formulas and examples of estimating sample size are presented. The correct interpretation of trial results is emphasized, and misinterpretations and errors that frequently occur are dealt with. Various issues regarding multiple significance testing, such as interim analyses, multiple endpoints, and subgroup analyses, are addressed.

The term *clinical trial* usually refers to any type of planned medical experiment involving patients. The rationale for performing a clinical trial is to determine the most appropriate treatment for patients with a certain medical condition. Results from a sample of patients are used to draw conclusions concerning the general population of present and future patients. Individual case studies and retrospective surveys are not considered as clinical trials.

The majority of clinical trials deal with various types of drug treatment and are often initiated by the pharmaceutical industry. However, other forms of treatment such as surgery, radiotherapy, and different forms of medical advice or patient care can be studied in clinical trials.

To document the effect of a new therapy in a scientifically acceptable way, a series of clinical trials has to be performed. The development of a new drug is divided into four different phases. Phase I trials usually include a small number of volunteers and are mainly concerned with pharmacological aspects such as safety and drug toxicity. In phase II small-scale studies of efficacy are performed, and suitable doses of the drug are identified. Phase III covers large trials comparing the effect of two or more different treatments. These trials are used to document the efficacy of a new treatment before applying for marketing authorization. Phase IV covers both long-term studies of adverse effects after approval, and trials for which marketing is the main objective. In what follows, the term 'clinical trial' will be used synonymously with a phase III trial.

The present paper will deal briefly with important aspects concerning design and analysis of clinical trials. The interested reader is referred to introductory textbooks on clinical trials (1) and practical statistics (2, 3). Two extensive papers on design (4) and analysis (5) of trials in which patients are observed over time are relevant for proper planning of cancer trials.

The CPMP Working Party on Efficacy of Medicinal Products (6) has published a European guideline to biostatistical methodology in clinical trials. Relevant principles of design and analysis are outlined, and generally acceptable approaches to these tasks are described. Some approaches which should not be adopted are also mentioned. Detailed specification of methodology is not given. The guideline can successfully be applied to both drug trials and clinical trials not involving drugs.

The trial protocol

Before conducting a clinical trial, a detailed protocol for the study must be written. All aspects of the trial should be described. The main features of a clinical trial protocol include the objective(s) of the study, study design, patient selection (inclusion and exclusion criteria), description of study treatments and clinical procedures, methods for eval-

From the Norwegian Cancer Society, The Norwegian Radium Hospital and University of Oslo.
Correspondence to: Prof. Eva Skovlund, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway.

uating patient response, randomization, blinding, sample size justification, statistical analysis methods, monitoring of trial progress, adverse events reporting, protocol violations and premature withdrawal, informed consent, and administrative responsibilities.

Trial design

Most trials are designed to include a certain number of patients which is decided in advance. This is called a fixed sample design. Sometimes, especially if the planned sample size is large, it may be of interest to analyse the data one or more times before reaching the planned sample size, and to stop the inclusion of new patients if a conclusion regarding treatment effect can be drawn. This is referred to as performing interim analyses. If the sample size is not fixed, but the data are analysed continuously during the trial, possibly after each patient responding, and the trial is stopped according to predefined stopping rules, it is called a sequential trial.

In order to assess the effect of a new therapy, it must in some way be compared with the effect of a standard therapy. Two main types of comparative design exist. A parallel group design implies that one group of patients receives a certain therapy and is compared with another group of patients receiving another type of treatment (the control group). The term cross-over design means that all patients receive both types of treatment during two different periods of time, and the order in which the treatments are given differs between patients. Patients are then said to act as their own controls, thereby reducing variability. With both types of design more than two treatments can be compared simultaneously. Cross-over trials generally need inclusion of fewer patients than trials employing a parallel group design, but their applicability is restricted by several assumptions, and they are for this reason not frequently used. The main restriction is that in order to ensure a fair comparison between two treatments given to the same patients in different periods of time, the disease under study must be chronic and stable, and the therapy to be studied must not be curative. Thus, a cross-over design is rarely feasible in cancer studies where a major endpoint is often survival or tumour response. The cross-over design should not be confused with a parallel group design in which patients in a trial are allowed to 'cross over' to the other study treatment after the treatment they were allocated to has failed.

Historical controls

A common way to avoid the inclusion of a control group is by comparing retrospectively patients receiving the new therapy with previously treated patients who received standard therapy. This latter group of patients is commonly referred to as *historical controls*. A major problem with historical controls is to ensure that the comparison is unbiased. If the groups of patients differ with respect to other characteristics than the therapy given, it cannot be guaranteed that an observed improvement is in effect due to the new therapy. Two major sources of bias exist; patient selection and change in experimental environment. Both types of bias tend to exaggerate the effect of a new therapy.

The historical control group will usually lack criteria for patient inclusion, and the investigator may be more restrictive in the choice of patients for the new therapy. As the controls were treated earlier, the type of patients available may also have changed. The quality of recorded data for historical controls tends to be inferior since these patients were originally not included in a trial, and retrospective collection of information will usually not provide the data needed. It may be difficult to ensure that the criteria for evaluation of response are the same in the two groups, and some aspects of patient management may have changed.

Randomization

In order to ensure comparability between treatment groups, patients should be randomized to receive one of the treatments to be studied. Simple randomization can be performed, for instance, by tossing a coin, but usually randomization lists are generated by a computer. To avoid selection bias, it is important that randomization of a patient does not take place until after he or she has been included in the trial. The purpose of randomization is to guard against systematic differences between treatment groups (except the treatments given in the trial). Simple randomization does not necessarily result in patient characteristics being distributed similarly in each treatment group, but any differences will be due to chance, and not to systematic arrangements made. Hence, standard statistical methods such as significance tests are applicable. Even if treatment groups by chance do differ considerably, for instance on the distribution of age or the proportion of women, the analysis is not invalidated. Analysis methods which allow for lack of comparability exist. However, scientists and readers tend to be more comfortable with treatment groups showing similar patient characteristics. To achieve similar groups, stratified randomization is often used.

Stratification

A stratified design is best illustrated by an example. When planning the trial a small number of patient characteristics believed to be essential for treatment outcome must be identified. In a breast cancer trial possible prognostic factors might be oestrogen receptor level ($+/-$) and tumour size (T1/T2). Stratification on these two factors, each on two levels, yields four (2×2) different strata

or subgroups. Within each stratum patients are randomized in blocks in order to ensure balance between the treatments, A and B, say. A common choice of block size is four; in which case six different permutations of two As and two Bs exist: AABB, ABAB, ABBA, BBAA, BABA, BAAB. Balance between treatments is thus ensured for every fourth patient included in the stratum. This concept is known as *random permuted blocks within strata.*

A drawback of this method is that if the block size is known to the investigator, it may, depending on the permutation of treatments within a block, be obvious at least for the last patient included in each block, to which treatment group a new patient will be allocated. Thus the investigator is not necessarily blind to which treatment the next patient will receive before including him or her in the trial, and this may lead to bias due to patient selection.

One should avoid stratification on a large number of prognostic factors since the number of strata will increase rapidly with the number of factors. By adding a third prognostic factor, e.g. N stage on three levels (0, 1–3, 4+), to the two described above, the number of strata would increase from four to twelve ($2 \times 2 \times 3$). The problem with a large number of strata is that some of them may remain empty or include very few patients, thus potentially jeopardizing the idea of balance between treatments.

With an increasing number of strata, balance between treatments within each stratum becomes irrelevant. The interest will instead lie on ensuring that the proportion of patients with a given characteristic is similar in both treatment groups. In statistical terms this is referred to as balancing the marginal totals. This may be achieved by using the *minimization* method developed by Pocock & Simon (7). A short practical description of the method is given by Pocock (1) and further details are described by White & Freedman (8). Instead of preparing a randomization list in advance, an updated record of treatment assignments by patient characteristics is kept, and the allocation of a new patient to a treatment group depends on his characteristics and on the characteristics of previous patients included. Table 1 illustrates the minimization method in a trial on adjuvant treatment of colorectal cancer. For simplicity, only two prognostic factors (Dukes' stage and localization of tumour) are included in the example. So far 84 patients have been included. If the next patient to be included has colon cancer, Dukes' B, he will be allocated to the surgery group, as sum surgery = $26 + 30 = 56$ and sum surgery + chemotherapy = $26 + 31 = 57$. There is not necessarily any true randomization, but an element of chance is usually introduced by assigning the adequate treatment with probability $p = 3/4$ or $p = 2/3$. In the example in Table 1 the new patient would be randomized to surgery with $p = 3/4$. The minimization method is frequently used in cancer trials (9).

In multicentre trials the centres entering patients may be considered as a factor for stratification in addition to relevant prognostic factors. Different hospitals may show different response rates due to patient selection and variability in experimental environment. With many centres the use of randomized blocks within strata will often result in an absurdly large number of strata. By using the minimization method the centre can be included as a factor in the same way as other patient characteristics.

Table 1

The minimization method in a trial on adjuvant treatment of colorectal cancer

Prognostic factor	Surgery	Surgery + chemotherapy	Total
Dukes' B	26	26	52
Dukes' C	15	17	32
Colon	30	31	61
Rectum	11	12	23
Total	41	43	84

Stratified randomization may require relatively large organizational resources. In very large trials stratification is unimportant (4). When patient characteristics are not readily obtained or there is uncertainty about the relevance of possible prognostic factors, stratification should be avoided.

Blinding

Even when patients are randomized between treatment groups the comparison between groups may be distorted if the patient or the physician knows which treatment is given. Ideally, randomized trials should be made double-blind, i.e. neither the patient nor the physician should know which treatment the patient is receiving. If no standard treatment exists, identically appearing placebo treatment by some inert substance should be given. If the standard therapy is unsatisfactory, as is often the case in cancer trials, both patients and physicians tend to be too enthusiastic about the new therapy, and knowledge of the study therapy given might influence patient response.

It is often impracticable to blind clinical trials in cancer research. In the case of comparison of chemotherapy regimens, knowledge of treatment allocation is often necessary due to toxicity. Dosage regimens may be very different, and the use of double dummy techniques, where patients receive either treatment A and placebo for treatment B or vice versa, may not be possible. Adverse effects may be treatment specific and easily recognizable such that blinding would almost inevitably be broken during therapy even if an attempt to perform a double blind trial were made. Radiotherapy trials might in theory be blinded by giving 'placebo radiation', but due to costs and the ease of breaking the blind, this is not done. Therefore most cancer trials suffer from not being blinded, and bias due to psychological effects on the patient and in the evaluation of response may be inevitable.

Bias due to psychological effects is not readily overcome, but the information given by the physician before randomization should be expressed in a way that minimizes this bias. One way to avoid bias in the evaluation of patient response is by keeping the evaluator blinded to treatment. The physician evaluating response should thus not be part of the treatment team.

Significance testing

The purpose of calculating a p-value is to assess the probability that an observed difference in treatment effect is due to chance only. A small p-value implies that the risk of a false positive result (i.e. to conclude that two treatments differ when they are in fact equal) is small. A real difference between treatments is then believed to exist. In clinical trials significance tests are usually two-sided, which means that if the p-value is small, it may be concluded either that treatment A is better than B or that B is better than A. If a treatment difference in one direction only (e.g. A better than B, but not vice versa) would be of clinical interest, a one-sided significance test may be performed. Whether to perform a one-sided or a two-sided test must be decided before starting the trial. Two-sided tests are usually regarded as a 'gold standard'.

It is customary to regard p-values lower than 5% as *statistically significant*, and the conclusion following the significance test will then be that one treatment is in fact better than the other. On the other hand, a p-value larger than 5% does not prove that two treatments have equal effect. Two treatments may actually be different, but this might not necessarily be detected by the analysis of patient responses. The probability of uncovering a true treatment difference is called the power of the test and should ideally be high. The power depends on the number of patients included. To achieve a high power it is necessary to include a large number of patients.

Sample size

Power based

Before conducting a trial it is very important to estimate the sample size needed to uncover a difference in treatment effect which would be of *clinical* importance to detect. The estimation of sample size can be illustrated by an example. Let the 3-year survival on standard treatment be $p_s = 40\%$, and let the clinically relevant improvement on the new treatment be 10%. The 3-year survival on the new treatment is thus hoped to be at least $p_n = 50\%$. A simple and useful formula for estimating sample size when comparing two proportions is

$$n = c \cdot (p_n(100 - p_n) + p_s(100 - p_s))/(p_n - p_s)^2$$

in each group. The constant c depends on the significance level and power chosen. If the significance level is 5% (two-sided) and the power to detect a given difference is 80%, then $c = 7.9$. Other useful values of c for relevant combinations of significance level and power have been tabulated (1). The constant c can easily be calculated by approximation to the normal distribution.

With the 3-year survival proportions above, it would be necessary to include $n = 387$ patients in each group. The smaller the treatment difference to be detected, the larger the sample size. It is clear that to detect small differences in effect, as are usually expected in cancer clinical trials, very large trials are required. Small trials will have lower power to detect clinically important differences, and will therefore often be inconclusive. The conclusion that two treatments are not different (i.e. the difference is too small to be of clinical interest) is only valid if the power is high.

Table 2

Power corresponding to different treatment allocation ratios

Allocation ratio	Power (%)
1:1	95
2:1	92
4:1	82

With a given number of patients randomization into groups of equal size is more efficient than allocating a different number of patients to each group. Nevertheless, there may be reasons, for instance ethical or practical, for including more patients on one treatment. In Table 2 it is shown that the loss of power is relatively small if the imbalance is not too pronounced. In the table groups of equal size are assumed to yield power 95% against a given difference in treatment effect. The total number of patients is kept constant throughout.

A p-value only states the probability that an observed treatment difference is due to chance and does not contain information on the size of the difference. Huge trials will tend to be 'significant' even when the difference is negligible from a clinical point of view, whereas small trials will usually fail to detect even large differences. The p-value depends on the number of patients included in the trial, and should always be accompanied by estimates of treatment effects. When two treatments are compared, a confidence interval for the effect *difference* should be estimated. A 95% confidence interval contains the true, unknown effect with probability 95%, and can be regarded as a measure of uncertainty. The shorter the interval, the more precise the estimate.

Precision based

Instead of basing the sample size estimation on the power of a significance test, a precision-based estimate can be made. This type of estimate can, for instance, be used when planning an uncontrolled phase II trial assessing the

proportion of patients showing a complete or partial tumour response. One then decides what approximate length a 95% confidence interval should have. To obtain a precise estimate (i.e. a short confidence interval), the number of observations must be large. The following formula may be of use for an approximate estimate of sample size

$$n = p(100 - p)(z/d)^2$$

where p is the expected percentage responding, z is a percentile from the normal distribution (for a 95% confidence interval $z = 1.96$), and d is half the accepted length of the interval. For illustration, let the expected proportion of complete responders on a given treatment be approximately $p = 60\%$, and assume it is of interest to estimate the true proportion with $d = \pm 5\%$ (i.e. an interval length of 10%). Then $n = 370$ patients should be included in the study. If $d = \pm 10\%$ were regarded as acceptable, $n = 90$ patients would suffice. These examples show that a large number of observations is needed to achieve high precision in estimates of proportions.

Some common methods of analysis

Knowledge of a few elementary statistical methods is essential to researchers and other readers of medical literature. The majority of articles published in medical journals refer only to a common set of well-known methods of analysis, and the reader would benefit from understanding this restricted set of methods. Proportions of responders are preferably compared by the χ^2-test or Fisher's exact test. Responses on a continuous scale can be analysed by t-tests or non-parametric tests of the Wilcoxon type (10). Advice on when to use which type of test can be found in elementary textbooks (2, 3). A general rule of thumb is to use non-parametric tests if the sample size is small; less than 10 patients in each group, say. It should be kept in mind that parallel group trials in general lead to comparison of treatments by means of two-sample methods, whereas the use of paired tests is restricted to situations where patients act as their own control.

Clinical cancer research frequently deals with analysis of survival data. Such data are typically incomplete since trials usually do not last long enough to observe the time to relapse or death for all patients included. The existence of incomplete observations of time to an event is referred to as censoring, and with such data special types of analyses are required. To estimate survival probabilities, the Kaplan-Meier method is extensively used. It is meant for graphical presentation of survival curves. The log-rank test is used to compare two or more survival curves. Instead of pairwise comparison between more than two treatments, it is wise to perform a log-rank test which compares all treatment groups in a trial simultaneously.

It is often of interest to use more complex analysis methods allowing for the inclusion of prognostic factors, thereby improving the precision of the estimated difference between treatments. If the treatment groups differ with respect to one or more prognostic factors, adjustment for these factors in the analysis may alter both the size and the significance of the treatment difference. When responses are quantitative, multiple linear regression is used to adjust for prognostic factors. With a dichotomous response (yes/no), logistic regression may be used, and for survival data proportional hazards models (usually referred to as Cox regression) are applied (Table 3). Such multivariable methods are complex and based on assumptions that are often violated. It is recommended not to perform such analyses without consulting a statistician.

Table 3

Multivariable regression methods often used in the analysis of clinical trials

Dependent variable (response)	Method
Continuous	Multiple linear regression
Survival times	Proportional hazards model (Cox regression)
Dichotomous (yes/no) or categorical	Logistic regression

Which patients should be included in the analysis?

Ideally, no patients should be lost to follow-up during a clinical trial, and all patients should be included in the statistical analysis of response. In practice, however, patients withdraw for a variety of reasons, they fail to comply with the treatment regimen they are randomized to, or experience side effects which make it necessary to stop treatment. Unless it is planned how to treat such patients in the statistical analysis, serious bias may occur. If, for instance, patients who withdraw from one treatment because it is ineffective are not included in the analysis, the treatment effect will be exaggerated. No true best solution to the problem of missing data exists. It is important, however, to plan the handling of such problems in advance, and to include a description of intended strategies in the trial protocol.

Broadly speaking, two main types of populations to be analysed exist. A 'per-protocol population' consists of patients complying with treatment and not violating assumptions made in the protocol. Such a population may be used to estimate the 'true' effect of a treatment when used according to protocol. Recognizing that patients are in practice seldom acting according to idealized standards, a so-called 'intention-to-treat population' is often analysed. Here, all patients are included, whether or not they comply with the protocol for the therapy they were allocated to. This may be regarded as a pragmatic approach to assessing treatment effect. The purpose of

analysing responses on an intention-to-treat basis is to remain conservative, i.e. to underestimate rather than overestimate a treatment difference. When drawing conclusions regarding treatment differences, the emphasis should be put on results based on the latter type of analysis population. Further discussion on this issue is presented by Pocock (1) and the CPMP Working Party (6).

Multiple endpoints

In most clinical trials several different endpoints or measures of efficacy are registered. For cancer clinical trials a common endpoint is the proportion of patients with complete or partial tumour response, no change, or progressive disease, according to specific criteria. Other frequently used endpoints are overall survival, cancer-related survival, and time to progression. Difference in toxicity between chemotherapy regimens may be analysed, in addition to comparison of quality of life measures over time.

The main problem with including a large number of endpoints is that the use of separate significance tests for each endpoint will increase the risk of false positive findings. With a 5% significance level 1 in 20 significance tests will be expected to be significant even when treatments are truly equivalent. A simple solution to this is to multiply each p-value by the number of endpoints subjected to significance testing. However, this so-called Bonferroni correction is conservative and tends to overcorrect, especially if the endpoints are strongly associated with each other. Instead, it may be preferable to reduce the number of endpoints registered, or specify in advance one or two primary endpoints. Selection of endpoints with low p-values after completion of the trial is a 'dirty trick' that leads to serious overestimation of the difference between treatments. Instead of reducing the number of endpoints, it may sometimes be feasible to combine multiple endpoints into a common score, as in quality of life research, where several items are added up to give one single score. The creation of meaningful scores of this kind is not an easy task, and analysis results should be interpreted with caution.

Subgroup analyses

A question that frequently occurs during analysis is whether the difference in response between two treatments depends on certain patient characteristics. Dividing patients into different subgroups and comparing treatments within separate subgroups is tempting, but may lead to problems. Separate significance tests do not provide direct evidence that a patient characteristic affects the treatment difference. The result of an interaction test would be more valid. Another source of bias is that data may be broken down in several ways, and problems with the interpretation arise since a large number of p-values can be calculated, thus leading to an increased risk of false positive findings. In addition, subgroup analyses create problems regarding statistical power; subgroups tend to be too small to uncover relevant treatment differences. It may sometimes be useful to display results for different subgroups, but subgroup analyses should generally be regarded as information of secondary interest to the overall comparison of treatments and any findings should be interpreted with caution.

Interim analyses

During the course of a large clinical trial it will both for ethical and for economic reasons often be of interest to perform one or more interim analyses before the planned number of patients has been included. If a significant difference between treatments can be demonstrated, the trial will be stopped. Thus, fewer patients will be included in the inferior treatment group, and more patients can benefit from receiving the better therapy.

However, interim analyses must not be performed unless they are planned in advance. Table 4 shows the increase in overall significance level when repeated significance tests are performed on accumulating data (11). The table is based on a normally distributed response with known variance, but similar results are also obtained with other types of distribution. Throughout the table it is assumed that the inclusion of patients will be stopped if an interim analysis yields $p < 0.05$, otherwise the inclusion of patients will continue until the next analysis is performed. From the table it can be seen that if a maximum of five analyses is performed, the true significance level, i.e. the probability of erroneously concluding that two treatments differ significantly when they are in fact equal, will increase from 5% to 14%.

To avoid the problem of increasing this error probability, the nominal significance level for each interim analysis should be reduced in order to keep the overall significance level constant at 5%, say. The maximum number of interim analyses must be decided in advance. It is customary to perform equally spaced interim analyses, either for number of patients or time intervals. Table 5 shows nomi-

Table 4

The risk of erroneously concluding with a significant treatment difference corresponding to the number of repeated significance tests performed

Maximum number of tests	Overall significance level (%)
1	5
2	8
3	11
5	14
10	19

Table 5

Nominal significance level for each repeated significance test when the overall significance level is 5%

Maximum number of tests	Nominal level (%)
1	2.9
2	2.2
3	1.8
5	1.6
10	1.1

nal significance levels which may be used for repeated significance tests with an overall level of 5%. The method has been developed by Pocock (11) and is often referred to as group sequential analysis with fixed nominal level. Other methods which, instead, apply variable significance levels throughout the trial are sometimes used (12). The main idea of all interim analysis methods is to make it possible to stop trials early without inflating the p-value. Pure sequential methods differ from interim analyses in that the trial may be stopped early, not only when a significant difference is found, but also if it can be concluded that there is no evidence of a treatment difference.

Publication bias

It is a common error that statistical significance is considered definite proof of a real treatment difference. Authors tend selectively to report significant trial results, whereas non-significant results are not mentioned. Occasionally this may be done deliberately, but more often it is probably due to the author being unaware of the resulting bias. The fact that trials with positive findings are more likely to be published makes the situation worse. Positive trials also tend to receive more attention from both editors and readers. It has been argued that perhaps the majority of trial reports claiming a treatment difference are false positives (1).

Conclusions

In order to draw valid conclusions from clinical trials it is important to avoid common errors in trial design and interpretation of results. Sophisticated statistical analysis methods can never replace proper trial design. To be conclusive, trials should include a control group, patients should be randomized between treatments, and potential sources of bias should be identified and dealt with at the planning stage.

Before conducting a trial it is very important to estimate the sample size necessary to draw a valid conclusion. Small trials are rarely able to uncover even large differences in treatment effect, and it may be argued that it is unethical to include patients in trials that are in advance known almost certainly to be inconclusive. It must on the other hand be remembered that a statistically significant difference does not necessarily imply clinical relevance. A p-value gives limited information in itself, and should always be accompanied by estimates of treatment effect, including a confidence interval. Unplanned interim analyses, and the reporting of an extensive number of p-values from subgroup analyses or multiple endpoints should be avoided.

REFERENCES

1. Pocock SJ. Clinical trials. A practical approach. New York: Wiley, 1983.
2. Altman DG. Practical statistics for medical research. London: Chapman and Hall, 1991.
3. Bland M. An introduction to medical statistics. Oxford: Oxford University Press, 1987.
4. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: I. Introduction and design. Br J Cancer 1976; 34: 585–612.
5. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: II. Analysis and examples. Br J Cancer 1977; 35: 1–39.
6. CPMP Working Party on Efficacy of Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products. Brussels: European Commission, 1994.
7. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975; 31: 103–15.
8. White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. Br J Cancer 1978; 37: 849–57.
9. EORTC. A practical guide to EORTC studies. Brussels: EORTC, 1994.
10. Conover WJ. Practical nonparametric statistics. 2nd ed., New York: Wiley, 1980.
11. Pocock SJ. Size of cancer clinical trials and stopping rules. Br J Cancer 1978; 38: 757–66.
12. O'Brien PD, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549–56.