TEXTBOOK OF

MEDICAL ONCOLOGY

SECOND EDITION

Edited by

FRANCO CAVALLI, MD, FRCP
Professor and Director
Oncology Institute of Southern Switzerland
Ospedale San Giovanni
Bellinzona, Switzerland

HEINE H HANSEN, MD, FRCP
Professor of Medical Oncology
The Finsen Center
Rigshospitalet
Copenhagen, Denmark

STANLEY B KAYE, MD, FRCP
Professor of Medical Oncology
CRC Department of Medical Oncology
University of Glasgow
Glasgow, UK

MARTIN DUNITZ
London
2000

First published in the United Kingdom in 1997 by
Martin Dunitz Ltd
The Livery House
7–9 Pratt Street
London NW1 0AE

Second Edition 2000

A CIP catalogue record for this book is available from the British Library

ISBN 1-85317-825-X

Distributed in the United States by:
Blackwell Science Inc.
Commerce Place, 350 Main Street
Malden, MA 02148, USA
Tel: 1-800-215-1000

Distributed in Canada by:
Login Brothers Book Company
324 Salteaux Crescent
Winnipeg, Manitoba, R3J 3T2
Canada
Tel: 1-204-224-4068

Distributed in Brazil by:
Ernesto Reichmann Distribuidora de Livros, Ltda
Rua Coronel Marques 335, Tatuape 03440-000
São Paulo
Brazil

Composition by Wearset, Boldon, Tyne and Wear
Printed and bound in Spain by Grafos, S.A. Arte sobre papel

3 Principles of clinical trials

Elizabeth A Eisenhauer, Marco Bonetti, Richard D Gelber

INTRODUCTION

Clinical trials are experiments conducted in human subjects for the purpose of evaluating one or more therapeutic interventions. Unlike observational studies, treatments are initiated according to a prospective plan for their evaluation, and data are collected purposefully and prospectively. All aspects of the study rationale, objectives, design, treatments, data requirements, statistical justification, and analysis plan are detailed in the protocol document. While laboratory experiments and observational studies in cancer patients can generate knowledge about cancer behavior, biologic determinants of outcome, and hypotheses about therapeutic benefit, it is only through clinical trials that therapeutic interventions can be reliably assessed. Since a clinical trial is an experiment, it must begin with a hypothesis or question: What are the toxic and/or biologic effects of a treatment or intervention? What effects does an intervention have on rates of tumor response, or on time to relapse, progression or death? Does it have a beneficial or detrimental effect on quality of life? To be successful, a study must clearly define the question and outcome measure(s) of interest, utilize an appropriate design and sample size to address the question, and collect and analyze data according to prespecified criteria. Furthermore, studies in human subjects must comply with international standards of ethical review and conduct. Finally, at the conclusion of each trial, the interpretation of its results should be undertaken in the context of pre-existing knowledge in the particular clinical setting studied. This chapter will address all of these aspects of clinical trials in cancer patients.

INTERNATIONAL STANDARDS AND HARMONIZATION

Today, cancer research, both basic and clinical, takes place in an international arena. Results of a trial in Europe or Japan can influence practice and research questions in North America, and vice versa. For this process to be effective, those engaged in clinical cancer research have required a common 'language' to describe clinical trial outcomes. Thus, in the 1970s, the World Health Organization (WHO) initiated a series of international meetings to standardize reporting of cancer study results. The resulting recommendations led to the publication in 1981 of an article: 'Reporting results of cancer treatment'[1] and were also made available in a WHO handbook. Sections were included regarding

Contents

reporting of baseline patient data, treatment delivery, toxic effects (WHO toxicity criteria), objective response, time to event results such as progression and survival, and general guidelines on reporting trial results (e.g. numerators and denominators, maturity of data, definition of 'cure', and more).

The WHO recommendations were widely adopted, but with time were found to be wanting in certain areas. The toxicity criteria were too simple, not allowing for precise descriptions of many toxic effects of treatment. An initiative by the US National Cancer Institute (NCI) in 1982 led to the creation of 'Common Toxicity Criteria (CTC)', which have undergone several revisions to respond to the need for expanded toxicity categories. The Response Criteria, as defined in the WHO handbook, have also been adapted by many international research groups to accommodate their individual research needs. Unfortunately, this has led to non-comparable definitions for response in some situations – the precise circumstance that the WHO criteria set out to avoid. To address this, and in response to a widespread recognition that the imaging tools to assess tumor burden have become more sophisticated during the past two decades, an International Working Party has begun the task of standardizing and simplifying response criteria once again.

The *Guideline for Good Clinical Practice* is the product of an international collaboration of regulatory agencies to achieve harmonized standards and requirements for the registration of medicinal products internationally (International Conference on Harmonization, ICH). This effort, which is more broadly based than just cancer research (but which applies to cancer clinical trials of investigational agents), has been underway since 1989. Topics for guideline development include quality, safety, efficacy, and statistical considerations in clinical trials. As each topic is identified, several formal steps of discussion and development take place, which culminate in a final guideline recommended for adoption by government regulatory agencies. For example, the *Guideline for Good Clinical Practice* completed all the steps in the development process, and was ready for adoption into domestic regulations as of May 1996. The European Union agency, the CPMP, adopted it in July 1996. Not all countries move with the same speed in completing the adoption step, but these guidelines remain useful international references, since there is a commitment in Europe, Japan, the USA, and Canada to adopt them when they been completed.

ETHICAL CONSIDERATIONS

International standards have also evolved with respect to the ethical conduct of trials and the protection of human subjects in medical research. The Nuremberg Code (1947) and the Declaration of Helsinki (1964) defined principles to govern biomedical research and protect the rights, safety, and well-being of trial subjects. Specifically delineated were issues of informed consent and its voluntary nature, the need for scientific rigor in the rationale and design of the proposed investigation, balance in the potential risks and benefits of the investigation, and the obligation to truthfully report results. Other jurisdictions have elaborated on these principles, in some cases defining regulations governing the conduct of clinical research. The *Guideline for Good Clinical Practice* reaffirms the principles of the Declaration of Helsinki. It also details the process of ethical committee review of research, the obligations of the investigator and the sponsor in clinical trials,

the required elements of clinical research protocols, and the documents that must be in place prior to beginning a clinical trial.

The ethical issues associated with phase I cancer trials have received special attention because the study population is a particularly vulnerable one: comprising individuals with the disease for whom standard or curative therapy has been exhausted.[2] In this setting, it is the task of the investigators to assure that the process of consent does not overemphasize the potential benefits of an untested treatment but instead conveys in lay language the true goal of the phase I trial: administration of gradually increasing doses of drugs to the point of unacceptable toxicity (see below). Although often unstated, the goals of the patient are likely to be at variance with this. Thus there are challenges to ethical committees, educators, and investigators to be aware of this special area of concern, and to train physicians in the communication skills required to discuss the benefits and risks of enrolling patients on phase I studies.[3]

Ethical issues are also particularly relevant for randomized clinical trials in which the choice of treatment from among an acceptable set of options is determined by a chance mechanism.[4,5] A distinction is made between the physician investigator, whose primary responsibility is for the welfare of the patient, and the clinical scientist, who also has an interest in conducting clinical research. Some argue that these roles are in conflict when the care of individual patients is at stake, thus making randomized trials inherently unethical. Others argue that applying therapies without substantial evidence concerning their safety and effectiveness is unethical and that the most valid way to obtain unbiased evidence is from randomized studies. Investigators must achieve a state of equipoise regarding the risks and benefits they perceive for each individual patient they subject to a random allocation of treatment. The extent to which physicians can achieve such equipoise depends upon the degree of uncertainty they acknowledge concerning the correctness of their 'best bet' of appropriate treatment for the patient. A balance is required between the patient's individual benefit versus the value of the trial to science and to future patients. When in doubt, the concerns for the individual patient should take precedence over the concerns for the study. Ethical studies are also those that seek to minimize any extra effort required of the patients to participate. Respect for the patients as individuals is an important aspect of ethical clinical trials.

THE PROTOCOL DOCUMENT

A carefully conceived and executed protocol document is essential for conducting a high-quality clinical trial. The elements of the protocol document are listed in Table 3.1. This table provides an excellent summary of the most important aspects of clinical trial design, conduct, analysis, and report. Whatever is not included in the protocol document is left to the vagaries of individual interpretation. This may jeopardize the study quality. Unless the essential elements are clearly defined before the trial begins, the potential for obtaining biased and uninterpretable results exists. For example, bias can occur if the protocol does not contain enough information to describe the procedures to be followed if a patient has a toxic reaction requiring dose reduction (or the cessation of a therapy), or if a patient has a recurrence of the disease that requires changes in treatment. Regulatory authorities rely exclusively on the protocol as the fundamental record of the prospective plan for the

Table 3.1 Elements of the protocol document

Study schema	A pictorial summary of the essential elements of the study design. This should be simple and logical, and should reflect the study objectives.
Background and rationale	A description of current state of knowledge that justifies the planning and conduct of the trial.
Objectives	Few in number, and achievable with the proposed study design.
Patient population	A clearly defined and ideally homogeneous cohort. Eligibility and ineligibility criteria should be easily verifiable at the time of patient enrollment, and the rationale for the criteria should be consistent with the study objectives.
Treatment allocation	Randomization used to avoid bias in comparative trials, stratification to achieve prognostic factor balance or to prospectively define intended subgroup analyses. Discussion of placebo control group, crossover plan, etc.
Treatment	Description of treatment administration, schedule, duration, potential toxicities, dose modification criteria.
Follow-up procedures	Patient visit schedules, and clinical laboratory data to be obtained.
Endpoints	Standardization and quantification of criteria for the evaluation of treatment effects. Definition of primary and secondary endpoints for analysis.
Statistical considerations	Review and justification of study design, presentation of sample size determinations, description of monitoring policies and plans for interim analyses, and outline of data analysis plan.
Forms submission	Simple, efficient, and comparable schedules for all treatment arms. Discussion of special quality control procedures.
Informed consent	Description of study rationale and objectives, risks, benefits, alternative treatments (including no additional therapy), and statement of the right to withdraw in terms that can be understood by the patients.

clinical trial. A well-written protocol not only serves to improve the quality of study conduct, but should also provide much of the material required for a meaningful presentation of the results at the conclusion of the trial.

DATA MANAGEMENT AND QUALITY CONTROL

Good data management practices are so essential for conducting a high-quality trial that they should be the responsibility of trained data managers rather than being left to busy clinicians.[6] Assurance that informed consent was obtained, proper verification of eligibility criteria prior to patient enrollment, and recording of treatment information are the first steps. Data monitoring of studies and quality control audits to verify that data reported on the case reports forms agree with the information in the patients' medical records are fundamental aspects of the *Guideline for Good Clinical Practice*. Data fraud in biomedical research is likely to be extremely rare, but when it is discovered the loss in confidence in the entire clinical trial process can be devastating. Protocols must therefore be designed so that they are feasible to conduct with good quality control standards.

THE CLINICAL TRIAL PARADIGM

To understand the statistical principles for the design and analysis of clinical trials, it is important to distinguish between the observed outcomes obtained from the study and the effects that the treatment might produce in the larger target population. This distinction is illustrated in Figure 3.1. The

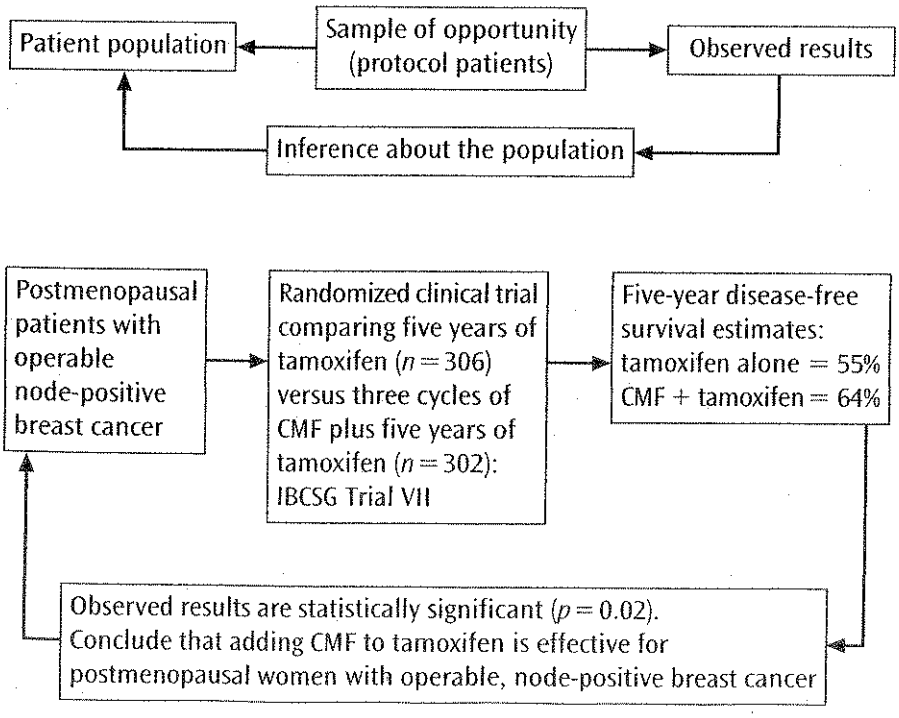

Figure 3.1
Clinical trial paradigm and example for a phase III trial.

protocol document defines the patient *population* of interest. The clinical trial recruits a *sample* of opportunity. The study endpoints obtained from the selected sample are recorded, and the results are summarized using statistical methods. The statistical methods are available to *describe* the observed results obtained in the sample and to *infer* what the true effect of treatment might be in the patient population based on the observed results. The lower portion of Figure 3.1 illustrates an example for a phase III study. In this case, the five-year disease-free survival percentages provide the descriptive estimates summarizing the study results, and the p values provides the statistical inference, indicating that the observed results could be expected only 2 times out of 100 if the two treatments were actually equally effective for the patient population.

The observed result from a clinical trial is the sum of the average true effect of treatment for the patient population, systematic error (or bias), and random error (or variability). The objectives of clinical trial methodology are:

- to minimize biases that might produce observed treatment effects that are distortions of the real impact of treatment;
- to minimize the variability of the trial estimates by evaluating an adequate sample size to detect treatment effects that might be reasonably expected.

All of the principles of clinical trials discussed below are designed to achieve one or both of these objectives.

ENDPOINTS

Standard descriptions of the endpoints of clinical trials are important if results are to be properly interpreted. The study endpoint is the information to be used to quantify the effect of the treatments being studied. It can consist of:

(a) a quantitative (discrete or continuous) measurement on each patient at a particular point in time, or 'longitudinally', i.e., several times throughout the follow-up time;
(b) the observation of whether or not a particular event has taken place; or
(c) the time elapsed between a pre-fixed time point (such as randomization or initiation of treatment) and the moment when a certain event occurs.

An example of a continuous outcome belonging to category (a) is the measurement of bone mineral density at the end of treatment. The emotional well-being of a patient assessed on a scale with multiple levels is a discrete quantitative measure with 'categorical' outcomes. The outcomes of type (b) are frequently defined endpoints in a clinical trial, and can be regarded as a particular ('binary') case of discrete outcomes. The occurrences of these outcomes are referred to as 'events'. Examples of events of interest can be disease progression within one year, response to treatment (such as tumor shrinking), or the occurrence of some toxic reaction.

The outcomes in category (c) are usually called 'survival times', but the final event that defines the time interval does not have to be death (in fact, disease-free survival is a very common endpoint). Not all patients will have had the event of interest by the time of the data analysis. The 'survival time' of these patients is only known up to a lower bound (the latest time that they were observed as not having had the event), and these 'survival times' are called 'censored'. The presence of

censored observations complicates the analysis of the trial data. In particular, the reported results may be subject to bias if the chance of censoring is related to treatment, for instance if some patients never returned for later visits ('lost to follow-up') because of toxic effects or worsening of their health.

Recently, there has been an increase in the use of surrogate measures of type (a) or (b), which may be correlated with survival, so that conclusions about treatment benefit can be made in a shorter time frame. In cancer clinical trials, an example of a surrogate measure is complete response to treatment, but complete responses occur too infrequently to provide meaningful comparisons. Surrogate measures can be used as the principal endpoint of a trial.

The choice of the endpoint of interest must be specified during the design phase of the trial, since it affects the selection of the statistical techniques for the analysis of the results, and the sample size required to properly evaluate treatments. Below, three types of outcomes frequently evaluated in cancer clinical trials are discussed: toxic effects, tumor response, and quality of life.

Toxic effects

Toxicity is an important measurement in cancer clinical trials, since most therapeutic interventions in cancer cause significant morbidity. This is related not only to the nature of the treatments employed, but also to the fact that treatment is generally delivered in as intensive a manner as possible. Toxicity criteria describe effects in categories or grades, with grade 0 meaning 'normal' or 'none' and grade 4 meaning life-threatening toxicity. The most commonly used criteria are the Common Toxicity Criteria developed by the US NCI. The original version contained 49 toxicity terms grouped into 18 categories. Modifications have added new terms, to categorize new effects of treatment and to allow for more comprehensive reporting. The 1998 Version 2.0 of the CTC contains over 200 toxicity terms grouped into 24 categories (available from web site http://ctep.info.nih.gov/CTC3/default.htm).

Toxicity terms include side-effects of treatment and/or disease (such as nausea, vomiting, edema, or bone pain), or other medical events (such as bleeding, infection, or ataxia), as well as objective findings (such as chemistry and hematology results). Most trials require a baseline assessment to document symptoms or residual toxic effects from previous treatment, and then periodic assessments of toxicity over the course of the therapy. In general, toxicity is reported in tabular form as the worst grade for each effect for each patient. Duration of toxic effects and relationship to study treatment may also be described. Finally, in randomized trials, comparison of toxic effects between treatment arms is an important component of the analysis.

Tumor response

Tumor response refers to the description of objective change in measured disease during the course of therapy. Criteria for response categories became widely used in the 1970s, and although many variations of the original WHO definitions now exist, most criteria include four different 'categories' that describe objective response to therapy. Virtually all criteria require a baseline assessment of measurable disease, and then the determination of the sum of the perpendicular products of bidimensional measurements, which is used as a reference point for comparison to the sums

generated by repeat measurements over the course of the study. Using the original WHO definitions, a complete response (CR) requires the disappearance of all evidence of disease. Partial response (PR) is a 50% or more decrease in the sums of products of bidimensional lesions, or a 50% or greater decrease in linear measurement of unidimensional lesions compared with baseline. Progressive disease (PD) is at least a 25% increase in the size of one or more lesions or the appearance of new disease. Stable disease (SD) (or no change) is a state that does not meet either PR or PD criteria. In addition, CR and PR must be confirmed by repeat measurements no less than four weeks after the first set of measurements were performed. In trials where objective response is an endpoint, patients are assigned a 'best response' to therapy on the basis of the above definitions. Clearly, to be evaluable for this endpoint, patients must have documented measurable disease at baseline and also have follow-up measurements during the study.

Despite these relatively clear definitions, there are frequent problems applying them in practice. For example, the methods for integrating the changes in size in unidimensional and bidimensional lesions vary among research groups when determining a 'best' response. Secondly, there are differences in how various researchers define the minimum size of lesions considered measurable, as well as the maximum number of lesions that should be recorded when multiple lesions exist. Some investigators assess response individually in each organ, while others consider the patient as a whole. Some investigators determine the duration of partial response from the onset of therapy, others from the date of 50% measurement change. Finally, many have altered the PD definition to be based on the sum of product increases rather than an increase in any one lesion. As a result of these problems, many variations of the original WHO criteria have evolved, leading to some confusion in the interpretation of the results of trials.[7] In addition, the application of different response criteria may lead to strikingly different conclusions about the efficacy of the same regimen.[8] Until uniform definitions of response are adopted worldwide, reports of clinical trials including a response endpoint should describe the criteria used.

What does achieving a partial response mean? Although an objective measurement of change in tumor size for a treatment group may signal that *the regimen under study has the potential to improve outcome*, achievement of partial response by an *individual patient* does not necessarily mean that he or she will experience a survival or quality-of-life benefit. Comparison of the survival of responders with that of non-responders is inappropriate, since responders must live long enough to respond, and often the very biologic factors that predict response are also predictive of longer survival. To attempt to address this issue, Buyse used a 'landmark method' of analysis[9] to show that, in colorectal cancer trials, response is a potent prognostic factor for survival in this disease.[10] It is interesting, however, to speculate that this may not be due to tumor regression, but rather that responders are segregated from the group destined to do poorly – whose 'best response' is progression. In a non-small cell lung cancer (NSCLC) randomized trial where response rates were low but survival improved over a non-treated control group, those patients who had partial responses and those whose disease was stable had identical survival outcomes.[11] These observations imply that survival may not always be improved by 50% shrinkage of tumor over and above the effect that disease stabilization may produce.

Unfortunately, the impact of tumor regression on symptom palliation has not been well studied. One study found that overall quality-of-life scores were improved or unchanged in a group of women who had either a partial response or stable disease in a trial in metastatic breast cancer, in contrast to those who progressed.[12] Surprisingly, continuous administration of chemotherapy was found to improve response rate, survival, and quality of life compared with intermittent administration for metastatic breast cancer.[13] Others studying palliative chemotherapy in colorectal cancer found that improvement in performance status or weight was associated with tumor response, but change in overall quality of life was not associated with tumor response.[14] More information of this nature would be useful, since greater or lesser degrees of tumor shrinkage that were *regularly* associated with symptom improvement might make achievement of a specific degree of measurable tumor regression a surrogate for palliation.

Quality of life

The quality of life (QoL) of the patient is an increasingly important endpoint for treatment comparison in clinical trials.[15] The US Food and Drug Administration (FDA) has stated that efficacy with respect to overall survival and/or improvement in QoL might provide the basis for drug approval.[16] The specific nature of the QoL endpoint to be used was not specified, and the statement has led to an almost automatic inclusion of QoL assessments in registrational trials. Including such assessments in clinical trials is relevant only if there is a clear hypothesis to be tested, and a methodology and analysis plan has been prospectively defined.[17] The objective of QoL assessment depends on the disease setting. In the adjuvant setting, toxic treatments are given to otherwise symptom-free patients with curative intent; assessments are to document the decreased QoL associated with treatment. For patients with advanced disease, however, QoL is measured to document the palliative benefits of therapies given to patients who are symptomatic. It is particularly important that QoL endpoints be justified, considering the extra burden placed on the patients to complete the QoL questionnaires.

QoL instruments have been developed that assess both global QoL and issues specifically related to cancer – the disease and its treatment. Some of the most frequently used are the European Organization for Research and Treatment of Cancer scale (EORTC QLQ-C30),[18] the Functional Assessment of Cancer Therapy (FACT),[19] the International Breast Cancer Study Group–Quality of Life Questionnaire (IBCSG-QL),[20] and the Quality of Life Index (QLI).[21] These instruments provide assessments of health status, describing patients' health-related QoL with respect to a variety of domains such as physical, psychological, and social functioning. Alternatively, QoL can be measured in terms of patient preference or utility assessments. Utilities are valued on a scale from zero (as bad as death) to one (perfect health), and reflect preferences for time spent in a particular health state. These assessments are useful for cost-effectiveness analysis and quality-adjusted survival. The Health Utilities Index (HUI)[22] and the Q-tility Index[23] provide both health status and utility assessments based on multi-attribute utility theory, which links utility values to measured health status.

QoL data can be analyzed using univariate, multivariate, or longitudinal methods. Univariate analyses are used to compare group

means based on individual scales. Multivariate analyses can adjust these mean values for multiple factors such as treatment, patient characteristics, and disease status. Longitudinal analyses evaluate change over time based on assessments taken at multiple time points. Carefully conceived analysis plans for QoL data are missing from most protocols. Statistical issues concerning how to handle drop-outs and missing QoL assessments must be considered.[24] For example, it is not clear in QoL-orientated treatment comparisons how to include patients who do not report QoL because they have died.

Quality-adjusted survival is an endpoint that considers both quality and quantity of life. Generally, this endpoint represents a patient's survival time weighted by the utility values representing the QoL experienced. One approach to evaluating quality-adjusted survival time within clinical trials is the Quality-adjusted Time Without Symptoms of disease or Toxicity of treatment method (Q-TWiST).[25,26] To apply the Q-TWiST method, clinical health states that can occur during the course of the patients' treatment and follow-up are defined to reflect changes in clinical status that may be associated with changes in QoL. For example, one clinical health state may be associated with toxicity due to treatment (Tox), another associated with disease progression (Prog), a third associated with development of late sequelae such as cardiac dysfunction (LS), and a fourth representing a relatively good state of health associated with none of the above (TWiST). The choice of clinical health states should be specific to the patient population and treatment comparison of interest, and should be designed to highlight QoL-oriented tradeoffs for treatment selection. The second step is to use a nested sequence of Kaplan–Meier curves to partition the overall survival time of each treatment group separately. The areas between these curves provide estimates of the average amount of time that patients spend in each of the defined clinical health states. The third step is to compare the overall quality-adjusted survival between the two treatment groups using a variety of weights for the health states. Sensitivity analyses based on threshold utility plots indicate the utility weights for which one treatment would be preferred to another. In cancer medicine, Q-TWiST has been used to evaluate adjuvant therapies for breast cancer,[27,28] interferon treatment for melanoma,[29] chemotherapy and radiation therapy for rectal cancer,[30] and chemotherapy for colon cancer.[31]

Cost-effectiveness analysis

There is an increasing emphasis on cost savings and cost containment in providing medical services and therapies. Consequently, economic evaluation has been included as part of many recently activated clinical trials. Methodological and statistical issues relating to this initiative were discussed at a recent symposium.[32] Economic evaluation refers to a comparative analysis of alternative lines of action, including their effects as well as their costs. Cost-effectiveness analysis (CEA) has become the most widely accepted method of economic evaluation of health care alternatives, according to the recent report by the US Panel on Cost-Effectiveness in Health and Medicine.[33] The CEA designed to evaluate one approach versus another yields a ratio of incremental costs divided by incremental effectiveness (measured in units of quality-adjusted life years, QALYs). Collecting individual patient cost data and measuring QALYs longitudinally represent challenges for the

conduct of economic evaluations alongside clinical trials.

ANALYSIS OF CLINICAL TRIAL RESULTS

The choice of the analysis method for the clinical trial data depends on the endpoint to be analyzed. The two most common measures of treatment effect encountered are the proportion of events and the survival distribution.

Proportion of events

The proportion of events is defined as the ratio between the number of patients in a group for whom an event has been observed and the total number of patients in that group. When computed for the sample of trial patients, this ratio is an estimate of the true proportion of events that one would observe if the treatment were to be assigned to the whole patient population represented by the sample. As an estimate based on a limited number of patients, the ratio would vary around the true population proportion if the trial were repeated many times.

The analysis of a trial whose endpoint is a proportion of events usually produces a *point estimate*, a *confidence interval*, and a statistical *test of hypothesis* about the population proportion. A point estimate is the observed proportion of successes, computed within each patient subgroup in the trial (e.g., within each treatment arm). A confidence interval for the population proportion of successes is a pair of values (p_1, p_2) such that the population proportion can be expected to fall within the interval with a given *level of confidence*. The precise interpretation of a confidence interval is that if one were to repeat the trial a large number of times (on different samples of patients) and compute a 95% confidence interval for the population proportion from each of the trials, then approximately 95% of the resulting intervals would actually contain the *true* population proportion. Confidence intervals can be generated that have any level of confidence between zero and 100%. The width of the interval will become larger and larger as the required level of confidence increases, including the entire range between 0 and 1 in order to achieve a 100% level of confidence. This is obviously not very informative. A confidence level of 95% is standard practice.

Confidence intervals can be constructed for a proportion, or for the difference between two proportions in a comparative study. Charts are available to readily obtain confidence intervals for proportions, and most statistical software packages will generate them.

A test of hypothesis consists of a statistical procedure designed to give an answer to questions of the kind: 'Can we reject the hypothesis (*null* hypothesis) that the two proportions of events p_A and p_B for treatments A and B applied to the whole patient population are identical?' Other questions that may be answered are of the kind: 'Can we reject the hypothesis that the proportion of events in a specific treatment arm is at least 20%?' The statistical theory of hypothesis testing in clinical trials involves specifying an *alternative hypothesis* and an admissible *type I error* (or α *level*). The type I error is the probability of rejecting the null hypothesis when it was actually true. A type I error of 5% is commonly accepted. The definition of the alternative hypothesis involves choosing the direction of the treatment difference to be considered 'extreme' in the testing procedure. Suppose that response to treatment is taken as the 'event' of interest, and that p_E and p_C are the response rates corresponding to the

Table 3.2 Analysis of proportion of responses in ECOG trial E3690 comparing dacarbazine plus interferon (with or without tamoxifen) versus dacarbazine without interferon (with or without tamoxifen) for patients with advanced malignant melanoma

	Complete responses	No complete responses	Total number of patients
Interferon	9	113	122
No interferon	4	124	128

experimental therapy and the control therapy, respectively. If the possibility that $p_E < p_C$ is not at all clinically interesting, then $p_E > p_C$ would be the alternative hypothesis of interest. Such an alternative hypothesis is called 'one-sided' to distinguish it from the 'two-sided' alternative hypothesis $p_E \neq p_C$, which will reject the null hypothesis whenever the evidence from the trial is either in the direction of $p_E < p_C$ or in the direction $p_E > p_C$. In the context of phase III clinical trials, the use of two-sided alternative hypotheses is highly recommended, since it is possible that a promising experimental therapy could actually prove to be *less* effective than the standard therapy.

The practical execution of a test of hypothesis requires consulting special tables, or examining the output from a statistical computer program. The test is said to *reject the null hypothesis at the level of significance* $100(1-\alpha)$ if the value of a quantity computed from the data (the 'test statistic') falls within a given interval. Such intervals contain all the values of the test statistic that are considered to be 'too extreme', where being 'extreme' is defined by the specified alternative hypothesis and the level of significance.

An alternative and equivalent way of conducting a test of hypothesis is the computation of a *p-value*. A *p*-value is the probability that a value for the test statistic equal to or more extreme than the one observed can be obtained *if the null hypothesis is assumed to be true*. Rejection of the null hypothesis is then declared whenever the *p*-value is smaller than the pre-assigned value of the type I error α, usually set at 5%.

The results from a clinical trial in which the endpoint is the occurrence of an event can be described in a 2×2 contingency table as in Table 3.2. This shows the result of an Eastern Cooperative Oncology group (ECOG) study in which 9 of 122 patients with advanced melanoma had a complete response with dacarbazine plus interferon therapy compared with 4 of 128 patients who received dacarbazine without interferon.[34] The observed complete response rates were 7% versus 3%, but the *p*-value calculated using a two-sided Fisher exact test was 0.16. Although the observed results suggest that interferon might be associated with a higher complete response rate, the observed results could have happened by chance alone (i.e., the true response rates were really the same) 16 times out of 100. Thus, the difference is not statistically significant. If the counts in the contingency table are large enough (a rule of thumb is that they should all be greater than

or equal to 5), then an approximate test can be conducted by computing the *chi-square* test statistic. Again, tables or a computer program will provide the rejection interval or the *p*-value.

Survival data

When the endpoint is a survival time, the population characteristic of interest is the *survival distribution* rather than just the number of events. A survival distribution is the probability distribution of the variable 'survival time', and it quantifies the probability that a randomly selected patient survives (or, more generally, 'does not have an event') until a given time point. There is one such probability for each time point, and one can plot a survival curve (or probability of surviving) versus time.

Just as with the proportion of events, survival distributions obtained from an individual trial can only provide an *estimate* of the true population distribution. The estimated curve (the equivalent of the point estimate) will also vary around the real population distribution. Estimation of a population survival distribution from the clinical trial data is complicated by censored observations, because only partial information (such as 'alive up to a certain time') is available for some patients' time to event. Two methods are most often employed to estimate the survival distribution, namely the life-table method and the Kaplan–Meier method.[35] The life-table method is used when a large number of observations is available, while the Kaplan–Meier estimator generally is used when there is a smaller number of observations (as is common in most clinical trials). Both methods are valid if the chance that an observation is 'censored' does not depend on the time at which an event (endpoint) might have been observed. The basic difference between the two methods is that the first provides estimates within pre-specified time intervals, while the second generates estimates at each time point where an event occurs. Both methods are readily available in most statistical software packages. Confidence intervals for the survival probability at a given time point can be computed, as well as confidence intervals for the difference in survival probabilities. An example of a Kaplan–Meier estimate of a survival function is shown in Figure 3.2. The following are some key features to look for and related questions to ask when examining a survival plot:

- the units on the vertical axis (Does the probability of survival shown range from 0 to 1, or is the scale truncated?);
- the scale of the vertical axis (Are the survival curves drawn according to a logarithmic scale or a linear scale?);
- the units on the horizontal scale (Over what length of time does the survival curve extend?);
- the amount of follow-up time available for the sample used to plot the curves (How many patients remain at risk to be included in the analysis at the tails of the curves?).

The plots in Figure 3.2 show a linear vertical scale ranging from 0 to 100%, a horizontal scale extending to six years from date of randomization, and a reasonable number of patients at risk out to five years (the median follow-up for the trial was five years at the time of the data analysis cutoff).

Tests of hypothesis can be conducted on survival distributions. Equality of the survival distributions corresponding to two treatments is the hypothesis ordinarily tested in clinical trials having a survival endpoint. Such a test is

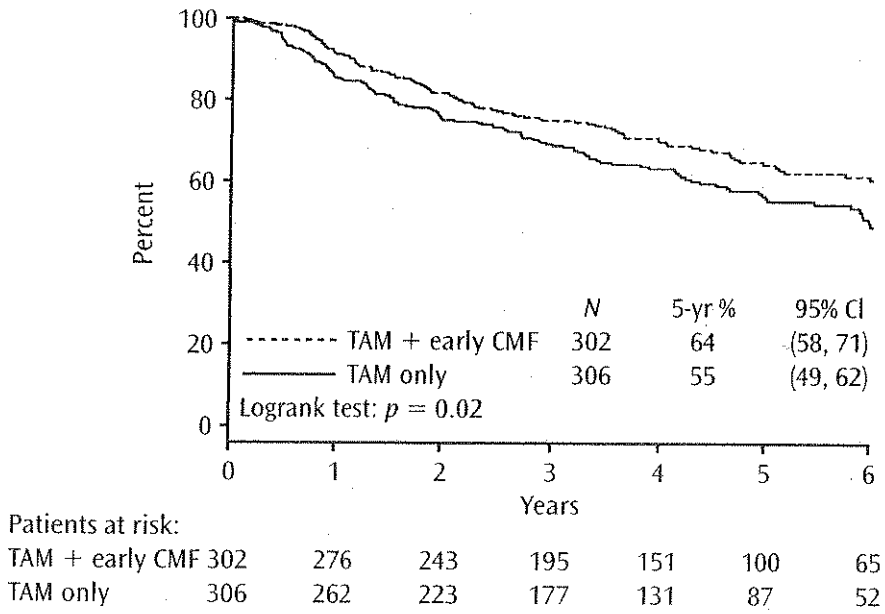

Figure 3.2
Kaplan–Meier estimates of the two disease-free survival distributions for tamoxifen plus three cycles of CMF (TAM + early CMF) and for tamoxifen alone (TAM only) from the International Breast Cancer Study Group Trial VII.[77] The median follow-up of patients for this analysis is five years. Also shown are the estimated percentages of patients surviving disease-free at five years from randomization, the 95% confidence intervals (CI) for these estimates, the logrank *p* values for the treatment comparison, and the number of patients who remain at risk over time. The hazards ratio estimate, comparing the risk of an event in the TAM + early CMF group with the risk of an event in the TAM only group, is 0.67 (95% CI 0.50–0.91), indicating a 33% reduction in the risk of an event with the combination treatment.

conducted by computing a test statistic called the *logrank* statistic. This test is most powerful under the assumption that the ratio of the logarithms of the two true survival distributions is constant over time – which is not true if, for instance, the two survival curves cross or there is a plateau in one or both of the curves. Cure rate models might be more appropriate for the latter situation. The mathematical expression of the logrank test statistic is complicated, but its computation is easily performed on computer systems.

Multiple regression analysis techniques are available to account for the effect of prognostic factors in the analysis of both proportions of events (logistic models[36]) and survival distributions (proportional hazards models[37]). These models are discussed in standard textbooks, and are also readily available in statistical packages.

DEVELOPMENT AND EVALUATION OF NEW THERAPEUTIC APPROACHES

Traditionally, clinical trials are divided into three stages or phases: phases I, II and III. Phase I and II are most easily differentiated from each other within the context of the development of a new agent or therapeutic maneuver, where clear goals are articulated

for each trial phase and decisions about the continued evaluation of the new agent are made at the conclusion of phase I and II studies. However, the concepts described here also apply to the evaluation of 'old' drugs in new combinations, although within this context trials may combine the goals of phase I and II studies within a single protocol. The principles in the design and analysis of phase III trials described below have identical application in the comparative study of new versus standard therapy and in the comparative evaluation of two or more standard therapeutic approaches.

Phase I trials

Goal and primary endpoint of phase I trials

Phase I trials are the first clinical studies of a new agent or new therapeutic approach. Because of the toxic nature of many cancer therapies, phase I trials are conducted in cancer patients who have failed standard therapy, rather than in normal volunteers. The *major goal* of these trials is to determine toxic effects and the recommended dose and schedule of the agent(s) under study for further evaluation. Secondary goals are the description of pharmacologic behavior, determination of target inhibition (if appropriate) and documentation of antitumor effects (i.e., response). In the case of *cytotoxic agents*, an assumption is made that the higher the dose, the greater the likelihood of efficacy. Because most such agents exhibit a dose–toxicity relationship, toxicity is regarded as a surrogate for efficacy: the highest safe dose is assumed to be the one most likely to be efficacious. Thus, toxicity is the *primary endpoint* of phase I trials of cytotoxic drugs, and is used to make recommendations regarding the dose of agent to be used in phase II evaluation. The phase I protocol defines in advance those toxic effects that, by their severity, limit further dose escalation (*dose-limiting toxicity, DLT*). In most studies, grade 4 hematopoietic effects of one (or several) days duration or grade 3 major organ toxicity are regarded as dose-limiting. The *maximum tolerated dose* (*MTD*) is defined as that dose producing a certain frequency of DLT within the treated patient population. *The recommended phase II dose* is one dose level below that producing an unacceptable frequency of DLT. It is important to note that there is some semantic confusion with these terms: in some protocols, the recommended phase II dose is referred to as the MTD.

Patient selection

Since phase I trials evaluate therapies of unknown efficacy and potentially serious toxicity in patients with a life-threatening disease, there has been much written on the ethical considerations and patient selection criteria for these studies.[38,39] In general, only those patients who have exhausted treatments that offer an improvement in survival or cure rates should be enrolled. Furthermore, when the route of elimination and toxic effects in humans are unknown, patients entering phase I trials should have normal or near-normal blood results for hepatic, renal, and hematologic parameters. Finally, the consent process should clearly outline the goals of the trial and the fact that both benefit and toxicity are unknown. As dose levels approach the recommended phase II dose (i.e., the dose thought most likely to be therapeutic), it is common practice to enroll 'better-risk' patients with prior treatment characteristics more typical of those who might receive the drug in phase II trials. In this way, the recommended dose will be assessed in a population similar to that in which it will be applied.

Phase I trial design

In broad terms, the design of a phase I trial involves treating small cohorts of patients at ever-increasing doses of the drug(s) until toxic effects limit further escalation. Doses generally are not increased within individual patients, in order to avoid confusion between acute and cumulative toxic effects, although some have challenged the need for this caution.[40] There are three elements that determine the duration of the trial and the total number of patients enrolled:

- the starting dose;
- the escalation method employed;
- the number of patients enrolled at each dose level;

Although the concern of investigators is to determine the recommended dose quickly and precisely, other considerations influence the design of phase I trials. These include both the desire to *minimize the risk* of exposing patients to unacceptable levels of toxicity and the desire *to treat as few patients as possible at low dose levels* that may be non-therapeutic.

Historically, the concerns for patient safety have been handled by utilizing a 'conventional' phase I design. Following completion of animal toxicology in two species (rodent and non-rodent, or two rodent species) and, in the absence of significant species differences, an accepted safe starting dose of one-tenth of the mouse equivalent LD_{10} (0.1 $MELD_{10}$) dose is calculated. If significant species differences are seen in toxicology, a lower starting dose is calculated: usually one-third of the lowest non-toxic dose in the most-sensitive species. Patients are accrued in groups of three, escalating the dose in each new cohort according to a modified Fibonacci sequence in which ever-higher escalation steps have ever-decreasing relative dose increments (e.g., dose increases of 100%, 65%, 50%, 40%, and 30–35% thereafter). The dose escalation is continued until the MTD is reached, usually defined as the dose producing DLT in 2 out of 3 patients or in 2 or 3 out of 6 patients. The next-lower dose level is the recommended phase II dose (RPTD). In practice, several schedules of administration for a new agent are taken into phase I trials. The final decision regarding schedule selection for phase II trials depends on a variety of factors, including toxicity patterns of each schedule, the total dose achievable, convenience, and preclinical data supporting schedule dependence.

This 'conventional' design is safe – seldom are patients given doses that are too high – but there are some problems with it, as follows:

1. If many dose levels are required to reach the MTD, large numbers of patients may receive non-toxic and thus presumably non-therapeutic doses. This raises an ethical concern, since, if any tumor responses are seen in phase I trials, they almost always occur at or near the MTD.[41]
2. The efficiency of this design is inconsistent. Although some trials may reach MTD with only 4 or 5 dose levels, in others multiple dose levels (10–15) are required, which may mean that the trial could last one or two years. With an increasing number of novel agents available, this type of time-line for completion of the first phase of investigation will delay the critical evaluation of these compounds in phase II and III trials.
3. The precision of the MTD estimate is limited because of small patient numbers at the highest doses. In addition, because phase I trials usually enroll patients with prior therapy, doses recommended within

this population may be lower than those easily tolerable within a population of untreated patients.

These problems have given rise to a number of different suggestions for altering the traditional, Fibonacci-based, phase I design. These have been reviewed at a large meeting on cancer drug development.[42] They include *increasing the starting dose*, enrolling *fewer patients per dose level*, especially in initial dose levels, *employing more rapid dose escalation schemes*, and *enrolling more patients at the recommended phase II dose level.*

Proposals for increasing the starting dose are based on a retrospective review of phase I trials, which suggested that, particularly for those agents in which two species' toxicology results were similar, higher starting doses (e.g. 0.2 $MELD_{10}$) would have been safe.[42,43] This suggestion has been made in conjunction with the traditional dose escalation scheme, but if more aggressive dose escalation is to be used (see below), the advantage of utilizing higher starting doses will be limited.

The enrollment of only one patient per dose level, particularly at the lowest doses of a phase I trial, has become a frequent and apparently safe approach.[44,45] This approach might be problematic if wide interpatient variability in toxicity is expected (such as for antifolates), when factors other than dose can contribute to the severity of toxicity.

Proposals to enroll one patient per dose level have often been combined with a proposed revision to the dose escalation scheme. Novel dose escalation schemes can roughly be divided into two categories.

The first of these is *empirically based* on doubling the dose in successive patient cohorts until an 'event' occurs that switches the escalation to a more conservative Fibonacci approach. The 'event' may be the observation of a toxic effect (accelerated titration design) or the achievement of a pharmacokinetic (PK) endpoint (e.g., 40% of the area under the curve (AUC) of the predicted MTD AUC based on animal pharmacologic studies).[46,47] Several groups have adopted these methods with apparent safety and increased efficiency. A recent publication used simulation techniques to show that accelerated titration escalation is safe with the accrual of one patient per dose level, switching to more conservative escalation once DLT in one patient or grade 2 toxicity in two patients had been seen.[40] Furthermore, it was concluded that this approach would save both patients and time in phase I studies. Similarly, pharmacologically guided dose escalation has been successfully applied to some trials but not others.[48] Limitations of this latter method relate to wide interpatient variability in PK seen with some compounds, problems with measuring or detecting blood levels, and the fact that in humans some compounds have active metabolites that are not seen in animals.

The second approach to more efficient dose escalation has been *statistically based.* Examples include the continual reassessment method (CRM) and the modified CRM,[49–51] model-guided determination of the MTD,[52] and escalation with overdose control (EWOC),[53] among others. These methodologies estimate MTD at the outset of the trial, and utilize toxicity data accumulated throughout the study to refine the hypothetical dose–toxicity curve and update the MTD estimate. Dose levels, which are often preassigned in 100% increments or some other fixed proportion, may be adjusted based on the adjusted MTD estimates. The goals of these methodologies are to provide rapid, safe

escalation to doses close to the MTD and to give a more precise estimate of the MTD during and at the end of the trial. Most practical experience has accumulated with the modified CRM method, which, in addition to the dose escalation process described above, usually involves enrollment of only one patient per dose level in the early stages of the trial. Two recent reviews suggest that the MCRM method is safe, treats more patients at or near the MTD, and enrolls fewer patients overall.[44,45] However, the method does not always appear to shorten the duration of phase I trials, probably because of the mandatory waiting period between dose levels to observe for drug-related toxicity.

Despite the plethora of new methods available for the design of phase I trials, a review has suggested that these are not commonly utilized by investigators.[54] This will probably change in the coming years as interest grows in expediting the phase I process. While most of the evidence suggests that enrollment of one patient per dose level and novel dose escalation schemes are safe and more efficient than the conventional Fibonnaci phase I design, little is known about how the new methodologies compare with each other, or about how well they perform in determining recommended phase II doses with precision. This latter assessment can only come from phase II and III studies of the new agent where the recommended dose is evaluated in larger patient populations.

Challenges in phase I trials of novel non-cytotoxic drugs

The discovery of novel agents targeting cell signaling pathways and the tumor cell microenvironment present challenges to phase I evaluation.[55,56] These drugs, as well as certain biologic therapies, may be active at doses that are non-toxic, thus eliminating the need to escalate to maximally tolerated doses. In fact, since some of these agents may be best administered in a chronic fashion, rather than in repeated high-dose cycles, it would be preferable if the doses selected for phase II evaluation did not have serious toxic effects. Thus, although the goal of the phase I study, namely to determine a recommended dose, is the same for non-cytotoxics, the endpoint may require modification. Direct measurement of target effect would be the ideal endpoint to substitute for toxicity, but there are practical problems with repeat tumor biopsies to allow measurement of target inhibition in the most relevant tissue. Alternatively, surrogates for activity, such as measures of modulated immune function for biologic therapies or blood levels of drug, may be appropriate. However, there is confusion about which measures are 'true' surrogates for efficacy. In fact, most trials of these agents continue to use toxicity as the parameter that limits dosing, while also measuring other parameters of target effect when possible to assist in final decisions about schedule and dose.

Phase II trials

Goal and primary endpoint of phase II trials

Once the recommended dose and schedule for a new therapy is available, phase II trials begin. These are small, uncontrolled studies in patients having the same tumor type. The primary goal of these trials is *to screen the new agent or regimen for efficacy and to estimate the level of activity*. In almost all instances, the complete + partial response rate (*overall response rate*) is the primary endpoint of such studies. Objective response is the only 'validated' endpoint for phase II trials: that is, it is

the only measure that has proven itself by allowing the identification of agents or regimens that are later shown to improve survival. While not all agents causing responses increase survival, some do. This endpoint presents a problem for phase II trials in some diseases, such as prostate cancer, where measurable disease is difficult to document. In this situation, other measures, such as change in tumor marker levels, are proposed to substitute for response in phase II trials. These measures still require validation. Furthermore, novel non-cytotoxic drugs may not be expected to cause tumor regression, so the use of objective response for phase II evaluation of these agents is problematic. More comments on this situation are to be found below.

Results of phase II studies of cytotoxic agents and hormonal therapy are used to make decisions regarding the future development of the agent or regimen. Although issues of patient selection and study design have common features for most phase II trials, it is useful to differentiate those studies that represent the first screening trial of a new single agent (phase IIA) from those that are non-randomized trials of drug combinations (phase IIB).

The single-agent phase IIA trial of a new drug determines if the experimental agent has a response rate above a targeted value deemed critical for pursuing the agent further. The target response rate is often set at 20%, but this may vary, depending on the tumor type and the patient population under evaluation. For example, a response rate as low as 15% might be of interest in a tumor where most drugs are inactive, such as melanoma, but a higher rate of 30% might be a more realistic target value in untreated breast cancer, which is responsive to numerous cytotoxic agents. Of major importance in this type of trial is the need to minimize the chance that a *truly* active agent is erroneously rejected. That is, the trial design should attempt to limit the probability of a false-negative result (type II error). False-positive conclusions about the activity of a new drug will be uncovered by its further evaluation following the initial phase II trial. However, if a false-negative conclusion leads to the rejection of the new agent, there may not be another trial conducted to provide the true information.

Phase IIB trials are similar in the sense that they are designed to determine the activity of a combination regimen against a target level of activity. For example, a new multidrug regimen in breast cancer might only proceed into phase III trials if in phase II it produces response rates superior to a prespecified target value determined on the basis of standard combination therapy results. However, other factors – primarily toxic effects – will be important to the decision about the regimen's future. Thus, in this setting, not only efficacy, but also toxicity, is estimated, and both parameters are important in the decision to pursue the regimen further.

Patient selection

Since response rates are the primary endpoint of phase II trials, it is important to recognize that factors other than the disease and drug under study will influence response and thus have an impact on the trial conclusions. For most malignancies, it is known that factors such as previous therapy, poor performance status, and large tumor burden can adversely affect response rates. Limiting adverse prognostic factors in the phase II population will stack the odds in favor of the agent showing activity, if such activity truly exists. An exception to this can be made in trials where an analogue or a resistance-modulating agent is

to be evaluated. In this situation, it may be of interest to enroll patients who have failed standard chemotherapy drug(s) in order to determine if the investigational agent(s) have activity in the setting of drug resistance. In addition to the general selection criteria noted above, other common entry requirements for phase II trials include adequate organ function (reflecting the fact that safe dosing guidelines for drug administration with impaired organ function are not usually available at this stage) and bidimensionally measurable disease. This latter criterion is essential for assessment of the response endpoint.

Whether to enroll previously untreated patients with metastatic disease into phase II trials depends upon the tumor type under evaluation and the efficacy of alternative therapy. For those metastatic or recurrent tumors in which no therapy has been shown to improve survival, such as melanoma, renal cell carcinoma, or malignant glioma, it is common to write protocols giving the investigational agent as the first-line therapy following incurable recurrence. For other diseases, such as metastatic breast cancer, untreated extensive small cell lung cancer (SCLC), NSCLC, and colorectal cancer, some investigators regard first-line investigational treatment as controversial.[57,58] It is argued that treatment with an investigational agent presents ethical concerns because it may delay active therapy and thus have an adverse effect on overall survival. Nevertheless, some groups have adopted the strategy of testing new drugs in these settings, usually under conditions of careful selection of patients and a 'switch' to active therapy in the absence of early response. A randomized trial of active versus (inactive) investigational therapy in SCLC provided some reassurance about the safety of this approach.[59]

Phase II trial design

As noted earlier, it is important in the design of phase II studies to limit the risk of false-negative conclusions and to provide an adequate estimate of activity in positive studies. The other major consideration that affects sample size and design is the desire to limit exposure of excessive numbers of patients to ineffective treatment. These competing requirements – to have enough patients to be sure the agent is truly inactive before it is rejected, but not so many that excessive numbers of patients receive inactive therapy – have led to phase II trial designs using two or more sequential stages of accrual. Should insufficient activity be seen after the first stage, the study will be terminated, only continuing to the end of the second (or third) stage if sufficient numbers of responding patients are noted.

A number of such multistage designs have been described.[60–63] The sample size for a phase II trial depends on the specification of a number of parameters that differ according to the tumor type, the patient population, and the drug under evaluation. These parameters include target levels of activity and the two desired error limits. The α error is a measure of the probability of accepting an inactive drug (false-positive result) and the β error a measure of the probability of rejecting a truly active drug (false-negative result). The multistage designs 'reject' an agent and stop the trial early if too few responses are seen at the end of the first stage. The numbers of consecutive non-responding patients that must be enrolled to reject a drug depend on the target response rate hypothesized to be of interest. The lower the response rate of interest, the greater the number of consecutive failures needed to conclude that the agent does not meet the hypothetical level of activity. Alter-

Table 3.3 Sample sizes for the initial cohort in phase II trials using the Gehan design

Target minimum response rate (%)	Number of patients in first stage if:	
	$\beta = 0.05$	$\beta = 0.10$
5	59	45
10	29	22
15	19	15
20	14	11
25	11	9
30	9	7
35	7	6
40	6	5
45	6	4
50	5	4

natively, to decrease the risk of false-negative conclusions (i.e., to decrease the type II error or β), the sample size in the first stage must be increased.

One example of a multistage design, the Gehan design, focuses primarily on elimination of drugs having a level of activity less than a prespecified amount, often 20%, which in practice translates into accrual of 14 patients and terminating the study if no responses are seen. This number is based on the fact that the probability of observing 0/14 responses when the true response rate is 20% is < 0.05 $[(1–0.20)^{14} < 0.05]$. If the agent passes the first stage, more patients are entered according to the level of presumed therapeutic effectiveness and the desired level of error (β). In the example given of 14 patients in the initial stage, the observation of 1 response will lead to an expansion of 11 additional patients to obtain a power of 90%, and an expansion of 45 additional patients to obtain a power of 95%. Table 3.3 shows examples of sample sizes for the initial cohort in phase II trials using the Gehan design. Depending on the target response rate of interest, and the β error, different numbers of patients are required to reject an agent if no responses are documented.

Both the Fleming[61] and the Simon[62] designs require that two 'target' levels of activity be hypothesized: a lower response rate (p_0) below which there would be no interest in pursuing the agent further, and a higher response rate (p_1) representing a level of activity of definite interest. In the Fleming design, the trial is terminated at the end of the first stage if extreme results are observed in either direction. That is, if the evidence supports either the hypothesis that the agent's activity is less than p_0 *or* that the regimen has definite activity above p_1, then the study is closed. It will only continue to the second stage if neither of these conditions is met. At the end of the second stage, recommendations to accept or reject the drug for further investiga-

Table 3.4 Examples of Fleming design and stopping rules for phase II trials

p_0 (%)	p_1 (%)	α	β	Number of patients 1st stage + 2nd stage = Total			First stage Reject ≤ resp	First stage Accept ≥ resp	End of study Reject ≤ resp	End of study Accept ≥ resp
5	20	0.10	0.08	15	20	35	0	3	3	4
5	25	0.09	0.09	15	10	25	1	3	2	3
10	25	0.11	0.10	20	20	40	1	5	6	7

tion are again made based on the final observed number of responses. Table 3.4 illustrates the Fleming design using as hypotheses a p_1 of 20% or 25% and a p_0 of 5% or 10% in various combinations.

A problem with the Fleming design is that, when the criteria for high activity are met, the trial is intended to stop early, since the answer to the question regarding level of efficacy has been addressed. In fact, in this circumstance, there are compelling reasons to continue accrual to allow enrollment of more patients. This not only broadens experience with the agent, but also provides better estimates of activity and narrower confidence intervals around the final observed response rate. In an attempt to address this, Simon[62] has published a method he terms 'optimal two-stage early rejection design'. This is similar to the Fleming design, but does not allow early termination when activity is seen. Instead, accrual continues to the second stage, where a final decision about interest in the agent is determined based on the number of responses seen in the final sample.

Other phase II design options deserve comment. The first of these is the concept of 'randomized phase II trials'.[64] Usually the term 'randomized' is used in the context of large-sample-size, phase III trials adequately powered to make comparisons between treatment arms (see the section below on phase III trials). However, in the phase II setting, there are some circumstances where it may be advantageous to randomize patients. The first instance is when two or more investigational regimens are available for study. Randomizing between regimens is a convenient way to accrue similar patient populations to simultaneous trials. In this case, no comparative intent is implied; rather, the technique of randomization provides an efficient means of conducting parallel studies, each with their own stopping points, etc. The second type of randomized phase II trial has been dubbed a 'pick the winner' design. Patients are randomized to two or more potentially active regimens, and the design is focussed on selecting the one most likely to be superior by a prespecified amount. Scenarios where this may be logical are when two doses or schedules of the same drug are pitted against each other to derive data allowing the selection of the 'winning' regimen to go into further study

against standard therapy. In this case, the trial is not adequately powered to convincingly determine the superior regimen; rather, the statistical calculation allows for reasonable certainty that if one of the regimens is truly more active by a prespecified degree (e.g., a 15% higher response rate), the probability will be high that the treatment with the greater true response rate will be declared the winner and selected for further study. This strategic approach has been used in randomized trials comparing various methods of administration of the cytotoxic agent topotecan.[65,66]

Another statistical approach of interest is one that may be utilized in phase IIB trials of combination-therapy regimens. In these trials, both efficacy and toxicity play roles in the decision to proceed further. In an effort to integrate these two endpoints in decision making, Thall has proposed a design whereby both adverse and efficacy outcomes are monitored on an ongoing basis.[67] For this design, a prespecified degree of toxicity is given, above which the trial would be discontinued. In addition, minimum efficacy requirements are outlined. In both cases, the levels for acceptance of the regimen are based on the efficacy and toxicity results of standard therapy. The trial may be designed, for example, to accept the experimental regimen if it produces 10% high response rates with no more than a 5% increase in serious toxic complications.

Other initiatives in phase II design include the examination of another endpoint besides response in a multivariate approach. Zee and colleagues have examined the utility of considering both response *and* early progression in decisions about early study termination.[68,69] This approach presumes that an agent with excessive rates of early progression will not have a great deal to contribute, even if it produces occasional responses. Both the number of early progressors and responders are evaluated at the early stopping point. The decision to continue to the second stage depends on observing both an adequate number of responses and fewer than a prespecified number of early progressions.

Interpretation of results of phase II trials

At the conclusion of a 'positive' phase II trial, the overall response rate should be reported along with the appropriate 95% confidence intervals. The denominator should include, as a minimum, all eligible patients who have had follow-up tumor measurements (i.e., evaluable patients). Some investigators report response rates as a proportion of all eligible patients regardless of their evaluability. These and other modifications to the types of patients included in the denominator can have a profound effect on the estimates of activity. Furthermore, phase II reports should include details about the population studied, especially with respect to those characteristics known to affect response. Finally, as some have shown, third-party review of films to verify responses will decrease response rates substantially.[70] Each of these elements will have an impact on the final response rate and thus the enthusiasm with which results are greeted. These factors, plus the varying criteria for response in use, also explain why widely differing results may be seen when the same agent is studied in apparently the same clinical setting.[71]

Challenges in phase II trials of novel non-cytotoxics

The discovery of novel molecular targets affecting cell signaling pathways and the tumor cell microenvironment present challenges to the standard phase II approach. Since in animal models many of these agents show tumor growth inhibition, but not tumor

regression, it can be argued that tumor response is not to be expected in clinical studies. These agents will need to prove their worth by prolonging survival, but, short of large randomized trials, are there other endpoints besides response that may screen these drugs for promising evidence of efficacy in small patient samples? A variety of alternative endpoints have been considered for this situation, including change in rate of rise of tumor marker levels, positron emission tomography, measurements of target inhibition, and clinical progression rates.[55,72] However, all of these alternative endpoints for phase II trials remain highly experimental: none have proven themselves by successfully selecting new drugs that have subsequently been shown to improve survival. Since many of these are currently being studied, we may soon have some evidence to indicate which might be valid for use in the screening of such compounds in phase II. Some investigators have chosen to avoid the problem of identifying the activity of non-cytotoxics in phase II by moving them directly from phase I to phase III trials. This approach is risky, but it may be justified if the preclinical data are compelling, if the results of the phase II studies will not change the plan to go on to phase III, and if interim analyses with early stopping points are designed into the phase III study.

Phase III trials

Goal and endpoint of phase III trials

Phase III clinical trials are conducted with the specific purpose of *comparing* one or more experimental therapies with the best standard therapy or competitive therapies. It should be noted that the best standard therapy may not exist, so that a placebo treatment may be the comparator. Endpoints for a phase III trial are selected to reflect comparative efficacy among treatments, and include time-to-event information, tumor response, toxicity, quality of life, or other measures of treatment effectiveness.

Patient selection

Since phase III trials are often the basis for widespread use of a new treatment approach, it is important to select a population of patients that is representative of the patients for whom the treatment will be used in practice. The patient risk profile should be such that the benefits are likely to outweigh the risks for study participants. Enrollment of patients with exceptionally good prognosis may dilute the ability of the trial to detect treatment differences, since the event rate will be lower than anticipated. Patients who might be at increased risk of suffering toxic effects of treatment (e.g., those with compromised hematologic, liver, renal, or cardiac function) should not be eligible for the trial. The eligibility and ineligibility criteria should be clearly described, so that the population to which the observed results might apply is well understood.

Generally speaking, a broad-based recruitment into phase III trials is preferred to a narrow set of eligibility criteria. This enables results to be extrapolated to a larger population, and provides the opportunity to study treatment effects across a variety of subpopulations. Recently, there has also been an emphasis on recruiting special populations – minority race and female sex – to clinical trials, so that treatment effects can also be evaluated within these groups.

Studies for pediatric malignancies

Conduct of clinical trials for the pediatric cancer patient population face special prob-

lems. Most pediatric malignancies are quite rare, making it difficult to recruit a sufficient number of patients to answer specific questions with much precision. Fortunately, clinical trial participation has become part of standard practice for pediatric oncologists, so that virtually every patient is offered a trial. The stakes are quite high for pediatric malignancies, where the opportunity for cure can provide many years and even decades of extended survival. These potential gains, however, must be balanced against the often-seen late sequelae associated with the disease and its treatment. Improved outcome and optimization of the therapeutic index for children with cancer during the past decades are the fruits borne by active evaluations of treatments within the context of controlled clinical trials.

Randomization, stratification, minimization, and binding

One of the main objectives of a phase III trial design is to eliminate systematic biases that could influence the comparison of the treatments. Bias can arise in many different ways, ranging from the selection process of the patients to be assigned to the different treatments, to the lack of proper follow-up, to the imbalance of the distribution of prognostic factors among the arms. One way to reduce bias is to make sure that a proper *randomization* procedure is used to generate comparable patient groups.[73] Randomization consists of the use of a chance mechanism to assign patients to the treatments, and it assures that each patient in a study has the same opportunity of being assigned to the therapies in the trial. It makes the treatment groups 'alike on the average' with respect to all factors (known or unknown) that may affect the endpoints of the trial. Neither the doctor nor the patient should know in advance which of the treatments will be assigned. Thus, randomization eliminates the possibility that the preferences of patients or clinicians will influence the constitution of the treatment groups. Simple randomization by chance can result in an unbalanced number of treatments being assigned. *Random permuted block randomization* assigns patients to treatments within blocks so that, after a specific number of assignments have been made (e.g., after every fourth or eighth treatment assignment), the number of patients assigned to each treatment is the same. *Unbalanced* randomization schemes can also be used in which the allocation of patients to treatments is unequal (e.g., two to one treatment assignment). A balanced assignment of patients to treatments is the most statistically efficient use of a fixed number of randomized patients. However, there might be practical issues or other advantages for using an unequal randomization – such as making a trial more acceptable by having a twofold chance of getting the experimental arm.

A strategy often used to further reduce potential bias is a *stratified randomization* plan. Patient subgroups (called *strata*) are defined with respect to a set of prognostic factors that may influence the outcome, and randomization is carried out separately within each stratum to guarantee balance among treatment groups. This feature is likely to be very helpful, especially in small trials or in the early stages of larger trials. A random permuted block design within strata can also assure prognostic factor balance across treatments within strata. It is important, however, to avoid including too many strata in the design, since this will reduce the trial's efficiency. More recently, a process known as *minimization* has been used to balance

treatment groups with respect to prognostic factors.[74] To use minimization, the prognostic factors of the patients in each treatment group are accounted for during the accrual period of the study. As each new patient is identified for enrollment, the sum of prognostic factors corresponding to those of the new patient is calculated for each treatment group. The patient is assigned to the treatment that would minimize the difference between prognostic factors in each treatment. If either treatment assignment would produce the same imbalance, then randomization is used to break the tie.

An additional safeguard against bias is the design of a *blinded* trial, i.e., one in which patients are not aware of the treatment to which they have been assigned. Such trials (when feasible) are termed *single-blinded*. A *double-blinded* trial is one in which neither the patient nor the physician know the treatment assignment. Blinded trials require that the physical aspect of the treatments be the same, and this is not always possible. Also, toxic effects clearly associated with one or more of the treatments may make blinding impossible. Blinding, however, is particularly valuable in trials for which the endpoint is subjective, such as pain relief. Blinding is also useful to improve compliance with treatment administration and to enhance the chances that the frequency and intensity of follow-up examinations are the same for both treatment groups. In addition, accurate assessment of the true extent of additional toxicity associated with treatment can only be assessed if the trial is blinded. *Placebo-controlled* trials use an inert substance in place of the active treatment in order to blind the study.

2 × 2 factorial, crossover, and equivalence studies

Several specialized types of designs are available to improve the efficiency of phase III clinical trials. *Two-by-two (2 × 2) factorial designs*[75,76] use all randomized patients to investigate two treatment interventions simultaneously within the same trial. For example, to evaluate the effects of treatments A and B, patients would be randomized to receive treatment A, treatment B, both, or neither. The effects of treatment A would be estimated by comparing the patients who received treatment A with those who did not receive treatment A, stratifying the analysis according to whether or not treatment B was also received. 2 × 2 factorial designs are most appropriate when there is no pre-existing knowledge suggesting a strong interaction between the effects of the two test treatments. An example of a 2 × 2 design is the International Breast Cancer Study Group Trial VII for postmenopausal, node-positive breast cancer patients.[77] In addition to adjuvant tamoxifen given to all women for five years, patients were randomized to receive three early cycles of CMF, three delayed single cycles of CMF, both early and delayed CMF, or no chemotherapy. The magnitude of the effect of early CMF and that of delayed CMF were each estimated using all of the patients enrolled.[77] In addition, the 2 × 2 factorial design allowed an investigation of the interaction of treatment effects. Neither of these advantages are present for a three-arm study designed to compare early CMF and delayed CMF separately against a tamoxifen-alone control group. 2 × 2 factorial designs are underutilized in phase III cancer clinical trials.

In *crossover designs*,[78] each patient receives each study treatment sequentially during different time periods. Patients are randomized to receive either treatment A during the first period followed by treatment B during the second period, or the reverse order of treat-

ment administration. The comparison of treatment effects is conducted within each patient, and the results are accumulated across all patients. This type of design is useful if a distinct endpoint can be measured within each time period to reflect the treatment effect obtained in that period. Crossover designs are very efficient because each patient provides his or her own control and the interpatient variability of the response is reduced. The presence of a period effect (difference in treatment response associated with the period of evaluation) or a period–treatment interaction (different treatment response associated with the period during which each treatment is received) will reduce the efficacy of the design and complicate the interpretation of the analysis. The pharmaceutical industry has favored these designs based on their efficiency, while the FDA has discouraged them because of the potential for confounding. The designs discussed above should not be confused with studies that allow patients who experience progressive disease while receiving one treatment to receive the alternative therapy at the time of progression. These studies might be more acceptable to doctors and patients, because patients who fail their first treatment have an opportunity to receive the second, but the crossover portion of the protocol is not intended to provide a direct comparison of treatments.

Another frequently used design is the *equivalence* trial.[79] In contrast to the usual *superiority* trial design, which seeks to provide evidence against the null hypothesis (to demonstrate superiority of one treatment over another), the equivalence trial seeks to 'prove' the null hypothesis by collecting evidence against the possibility that the test treatment is really worse than the standard. Because a clinical trial can never fully guarantee that the two treatments are exactly equal, these trials have recently been called '*non-inferiority*' trials, because the goal is to demonstrate that the new treatment is not inferior to the old. These designs are useful if the new treatment is less toxic, less costly, and more convenient than the old, and would be recommended on these grounds as long as it were shown to be equally effective. The sample sizes for equivalence trials are ordinarily larger than for superiority trials, because the magnitude of treatment inferiority that would be tolerable is usually very small. In addition, the stated inferiority threshold built into the study design influences the actual test of hypothesis for treatment differences; if someone disagrees with the choice of this quantity, the results of the study can change. Thus, equivalence trial deigns should be used sparingly, and only when the therapeutic question cannot be addressed by a superiority design.

Sample size determination for phase III trials

All quantities evaluated on the patients constitute *estimates* of the corresponding characteristics of the patient population, and as such they are subject to variability. Patients enrolled in a trial are only a (usually very small) portion of the potential population of patients that one has in mind when designing the study. If the same trial were to be conducted twice (on two different samples of patients), the numerical values of any measure aimed at quantifying treatment effect (such as the proportion of patients still alive after five years, or any other endpoint) computed in the two trials will in general not be *exactly* identical. The *precision* of an estimate defines how close it is likely to be to the true (unknown) population value. The precision of the estimate depends on the sample size of the trial.

The sample size determination depends on many features: the endpoint of interest, the goal of the data analysis, assumptions about the true variability of observations, specification of the magnitude of treatment differences of interest, and the size of acceptable statistical errors (type I and type II). For a time-to-event endpoint, the sample size is based on the number of events observed at the time of analysis; additional assumptions about the event rate, the accrual period, and the subsequent follow-up interval are thus required to determine the number of patients to enroll. Occasionally, sample sizes are based on the number of patients needed to obtain a 95% confidence interval for the difference between two treatments that is no wider than a specified amount. Most often, the sample size is determined such that the final test of hypothesis will have a prespecified *power* $(1 - \beta)$. The power of a test is the probability that the test will reject the null hypothesis if the true treatment effect difference is bigger than a prespecified clinically worthwhile amount (i.e., *when a specific alternative hypothesis is true*). A power of 80% or better is commonly used. Specifying the alternative hypothesis for the analysis of proportions involves guessing (usually on the basis of previous knowledge) the event rate for the traditional treatment, and specifying the smallest clinically relevant difference in event rate between such therapy and the new treatment under investigation. Together with the selected type I error of the test (traditionally 5%), these quantities determine the minimum of patients that must be recruited in each arm. Table 3.5 shows the required sample sizes per treatment group for a variety of situations.[80]

For survival analyses, the determination of the sample size is computed in a similar way. A survival distribution can be summarized by its median survival time, which is the time point at which 50% of the patients are still alive. Alternatively, the probability of survival at a given time point (e.g., at five years) can be used. A guess (again based on previous knowledge) must be specified for the median survival corresponding to the traditional therapy, as well as the smallest clinically relevant difference in median survival between the two treatments. The desired power of the test and its type I error must also be provided. Based on these quantities, the required number of patients who have an event of interest can be determined. A sufficient number of patients must be enrolled and followed for a long enough period of time to enable the required number of events to be observed. Studies in populations of patients at low risk of an event, therefore, require enrollment of many more patients than studies in high-risk populations. The anticipated event rate, the accrual rate, the accrual period, and the follow-up period are specified to determine the total number of patients required. Tables such as that shown in Table 3.6 are then available to determine the required sample size.[81,82]

Interim analysis, stopping boundaries, and data safety monitoring boards

One of the recent advances in the statistical design of randomized clinical trials has been the development of procedures that facilitate interim analysis of study data without increasing the risk of obtaining a false-positive result based on multiple looks at the data. Flexible group sequential methods are now available to allow early stopping of trials if interim results indicate either a striking advantage for one treatment over the other or no difference at all.[83–85] Software is also available to design sequential stopping boundaries.[86] A key

Table 3.5 Number of patients per treatment for a comparative trial based on proportions (assuming a two-sided false positive rate of 5% and equal allocation of patients to each treatment) [a]

Magnitude of clinically important differences between proportions			Sensitivity or power			
r_1 (%)		r_2 (%)	0.5	0.7	0.8	0.9
10	vs	15	375	579	725	957
10	vs	20	117	176	219	286
10	vs	30	40	58	71	92
10	vs	40	22	31	38	48
40	vs	45	791	1245	1573	2093
40	vs	50	210	324	407	538
40	vs	60	58	86	107	139
40	vs	70	27	40	48	62

[a] For more extensive tables, see reference 80 (Table A.3).

Table 3.6 Sample size for a survival endpoint, illustrating the role of accrual period and additional follow-up time: number of patients per treatment to detect an improvement in median survival from 12 months to 18 months (level of significance 5%, power 80%)

Years of accrual	Years of additional follow-up		
	1	2	3
1	150	117	104
2	132	110	103
3	122	107	102

feature of all such designs is that the characteristics of the design must be established prior to initiation of the trial and any data analysis. If boundaries are altered to fit the observed data, the protection against an inflated false-positive error rate is violated.

Many phase III clinical trials are currently conducted under the guidance of a Data Safety Monitoring Board. Principles defining the composition, independence, and responsibilities of this board have recently been described by the US NCI.[87] Essentially, interim results and safety monitoring data are periodically presented to the board (consisting of independent scientists and patient representatives), and the board advises the study investigators, concerning the continued viability and ethical conduct of the trial. In this way, investigators who are participating in the trial are not exposed to interim study results that might bias their participation. The committee is charged with protecting the interests of patients in the study (both current and future) and assuring that the study continues to meet the highest ethical standards for clinical trial conduct.

Challenges for the future

Increasing the number of cancer patients who enroll in phase III randomized clinical trials is an important challenge for the future. Moderate improvements in outcome for common diseases can have a substantial impact in terms of number of lives saved. Individual trials often do not enroll enough patients to provide reliable evidence to detect such moderate but important advances. The links between clinical trials research and bench sciences should continue to be strengthened. Randomized clinical trials must be conducted with prospective creation of a pathology tissue bank to be used to evaluate current and future markers that might predict response to treatment. Not only is it important to determine whether or not a treatment is effective, it is also important to determine those patients for whom the magnitude of the effect might be largest. Biological correlates of treatment effectiveness are likely to be extremely important for tailoring treatments for specific patient populations in the future.

Phase IV trials

The term 'phase IV clinical trial' usually refers to a clinical trial designed to monitor a new drug after it has been approved to be marketed. The goals of the monitoring usually relate to adverse effects as well as additional large-scale, long-term studies of morbidity and mortality. Another goal of these trials is the identification of subsets of patients for whom the drug performs particularly well. This can be very important for marketing purposes, as well as for further developments of the drug.

INTERPRETING THE RESULTS OF TRIALS

Guidelines for reporting clinical trials are designed to assure the reader as to the quality of the study and to improve the accurate interpretation of the results.[88] Recently, a consensus has been published on the consolidation for standards for reporting trials (CONSORT statement).[89] The guidelines presented in the consensus statement are used by many medical journals as the standard for reporting clinical trial results.

Intention-to-treat analysis

Randomization in phase III trials is intended to reduce systematic bias that could confound

treatment comparisons. Excluding patients from data analysis can, however, defeat the purpose of randomization and introduce bias. For example, in a trial comparing chemotherapy versus no chemotherapy, excluding patients who do not complete chemotherapy could leave only the 'healthiest' patients in the chemotherapy group, while all patients would be included in the no-chemotherapy group. Even if the trial were randomized, it is clear that the patients who remain after the exclusion are not comparable between the two groups. The *intention-to-treat principle* calls for including all patients in the primary data analysis according to their randomized treatment. The intention to use chemotherapy is compared against the intention not to use chemotherapy. Introducing systematic bias by patient exclusion is avoided. Although the treatment effect estimated from the trial could be diluted by including non-compliant patients, the potential for bias is a greater concern. Intention-to-treat analysis has, therefore, become standard for the primary evaluation in comparative clinical trials.

Interpreting the *p*-value and role of confidence intervals

The interpretation of *p*-values is often a source of confusion. The *p*-value resulting from a statistical test of hypothesis is *not* the probability that the two treatments are equivalent (or different), but rather a measure of the extent of the evidence against the null hypothesis. The distinction between statistical and clinical significance should also be clear: *p*-values have nothing to do with the quantification of *treatment effect*, which is measured by the difference in response rates, or in median survival between the treatment arms. To illustrate this point, consider two clinical trials, each designed to compare the effect of two treatments, and suppose that the results of the two trials can be summarized as follows:

	Observed response rate for arm A	Observed response rate for arm B	*p*-value
Trial 1	20%	30%	0.3
Trial 2	20%	30%	0.03

The conclusion that the two treatments compared in the first trial do not differ in effect, while the two treatments compared in the second trial do differ, is wrong. In fact, the large *p*-value in the first trial is due to the fact that the trial was not powerful enough to detect the difference in response rate, i.e., its sample size was probably too small. The point estimates and 95% confidence intervals for the difference in response rates provide a more informative summary of the estimated magnitude of the treatment effect difference. These estimates for arm B versus arm A are +10% (−10% to +30%) for trial 1 and +10% (+3% to +17%) for trial 2. In both studies, the observed response for treatment B is 10% higher than for treatment A, but the 95% confidence interval for the difference includes 0% in trial 1 (hence, the *p*-value is greater than 0.05) but does not include 0% in trial 2 (hence, the *p*-value is less than 0.05). The confidence interval provides a range of plausible values for the difference in response rate based on the observed data.

In phase III cancer clinical trials, the treatment effects realistically sought after can be quite small. However, even moderate benefits can be of real clinical importance

because of the large number of patients that a change in therapy might potentially affect. The detection of a small clinically relevant difference in response can require quite large sample sizes.

Subgroup analysis

The limitation of subgroup analyses is one of the more troublesome aspects of clinical trial analysis and interpretation. Evaluating treatment effects separately for multiple patient subgroups provides the opportunity to obtain more positive findings than could be anticipated by chance alone. Furthermore, if the overall result is statistically significant, there is a high probability that randomly selected subgroups of patients will produce strikingly different outcomes just by chance alone. For example, if the overall observed size of the treatment effect is two standard deviations, there is a one in three chance that one randomly selected subset of half of the patients will produce an observed effect of three standard deviations (highly statistically significant) while the other half will produce an observed effect of only one standard deviation (not at all statistically significant).[90] Therefore, when such effects are found in the data, we might report the findings, but not believe them.

The results from subgroup analyses, however, more closely meet the clinical objective to tailor treatment for individual patients. Unfortunately, the problems cited above are likely to make the overall results of the trial more representative of the true treatment effects within subgroups than the results of the subgroup analyses themselves. Thus, the results from the overall study should remain the major component of the report on a treatment advance. Subgroup analyses can be used to generate hypotheses that require confirmation in future studies. Subgroup analyses can be considered with less concern if there is a prospectively defined hypothesis of a differential treatment effect to be tested, as evidenced by a prospective stratification of the randomization.

Meta-analysis

Meta-analysis refers to the process of performing an analysis of the combined results of several separate studies.[91] The main purpose of meta-analysis is to increase the statistical power to detect treatment effects. Unfortunately, the magnitude of treatment effects obtained from currently available cancer therapies (especially in terms of improving overall survival) is relatively small. Therefore, only large trials have a good chance of achieving statistical significance. For example, over 3000 patients followed for an average of five years are required in order to have an 80% power of achieving statistical significance if the true effect of treatment is to increase five-year overall survival from 50% to 55%. The lack of statistical significance in several undersized studies creates the impression that the treatment is ineffective. Controversy arises if one or two trials achieve statistical significance despite the apparent overwhelming evidence from other trials to the contrary. A meta-analysis is conducted to distinguish modest but real treatment effects from the play of chance.

A meta-analysis proceeds in four steps. First, the research question is defined. For example, what is the evidence that adjuvant tamoxifen improves survival for patients with operable breast cancer? Second, a literature review and contact with investigators in the field are carried out to identify all of the studies available to answer the question. Because there

is a higher likelihood for positive studies to be published, relying only on published reports will tend to introduce a positive publication bias into the meta-analysis.[92] Third, the specific studies to be combine are selected and the data obtained. Such data might be derived as summary measures from published reports or from investigators in the form of individual patient data (IPD). The IPD approach is preferred, because unpublished studies can be included, some data quality control checks can be performed, and greater flexibility in the analysis can be achieved. Fourth, the statistical analysis to combine the separate studies is carried out, and the results are interpreted.

While increasing the statistical power by combining results from several studies, it is critically important to reduce the chances that systematic bias may be introduced in the results. The term 'overview' has been suggested to describe a specific type of meta-analysis conducted using strict guidelines to reduce such systematic bias. Overviews are based exclusively on properly randomized clinical trials. All available randomized trials (both published and unpublished) that investigate the therapeutic question of interest should be included. All patients should be evaluated according to their randomized treatment assignment (intention-to-treat analysis). Estimates of treatment effects based on comparing like patients within each individual study are first obtained, and then statistical methods are used to combine these separate treatment effect estimates into an overview result.

Groups of investigators have collaborated to perform meta-analyses on a wide variety of therapeutic issues in oncology. The Early Breast Cancer Trialists' Collaborative Group (EBCTCG) coordinated by the Oxford University Clinical Trials Service Unit has conducted a series of ambitious and instructive overview meta-analyses evaluating chemotherapy, tamoxifen, ovarian ablation, and radiation therapy for breast cancer.[93–95] One of the primary questions was whether adjuvant therapy with tamoxifen improved survival (reduced the risk of death) for patients with operable breast cancer.

Despite the substantial increase in statistical power, meta-analysis cannot answer all questions concerning the worth of a given treatment. In particular, the overall estimates of the magnitude of treatment effect obtained from the entire cohort of patients may be misleading. The magnitudes of the treatment effects estimated in the overview are based on an 'arithmetic construction' that ignores quantitative interactions between the effect of the test treatment (e.g., tamoxifen) and the patient characteristics or concomitant treatments (e.g., estrogen receptor status or chemotherapy).[96] In addition, indirect comparisons of treatment effects across different subgroups or across different overviews reflect the nature of the trials that were conducted. For example, an indirect comparison of the effect of tamoxifen in trials that used five years of treatment versus trials that used one year of treatment is not as valid as a direct randomized comparison of five versus one year.

Meta-analysis of cancer clinical trials is a very powerful procedure that stimulates international collaboration, helps to resolve controversies that arise from apparently contradictory results of undersized trials, and focusses attention on major questions that remain unanswered. Nevertheless, well-conducted, large-scale randomized clinical trials continue to be the basis for evaluating the worth of new therapies and regimens for cancer patients.

Can clinical trials be the treatment of choice for patients with cancer?

One principle of clinical trials of cancer therapies is that the studies are only as good as the data that go into them. A feature that has substantially slowed progress in improving the use of current therapies and developing new treatment approaches has been the disappointingly low percentage of patients who participate in randomized clinical trials. Several studies have been conducted to explore reasons for this low participation.[97–99] Because there are very few cancers for which an adequate cure rate can be achieved with current modalities, future advances are still required. Efforts must be made to make participation in clinical trials more widely acceptable. Patient advocacy groups are recognizing the important role played by a vibrant cancer clinical research effort. More progress can be made only if clinical trials become the treatment of choice for many more patients with cancer.[100]

REFERENCES

1. Miller AB, Hoogstraten B, Staquet M, Winkler A, Reporting results of cancer treatment. *Cancer* 1981; 47: 107–14.
2. Sessa C, Cavalli F, Who benefits from phase I studies? *Ann Oncol* 1990; 1: 164–5.
3. Daugherty CK, Ratain MJ, Siegler M, Pushing the envelope: informed consent in phase I trials. *Ann Oncol* 1995; 6: 321–3.
4. Hellman S, Hellman D, Of mice but not men, problems of the randomized clinical trial. *N Engl J Med* 1991; 324: 1585–9.
5. Passamani E, Clinical trials, are they ethical? *N Engl J Med* 1991; 324: 1589–92.
6. McFadden E, *Management of Data in Clinical Trials*. Wiley: New York, 1998.
7. Tonkin K, Trichler D, Tannock I, Criteria of tumor response used in clinical trials of chemotherapy. *J Clin Oncol* 1985; 3: 870–5.
8. Baar J, Tannock I, Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. *J Clin Oncol* 1989; 7: 969–78.
9. Anderson JR, Cain KC, Gelber RD, Analysis of survival by tumor response category. *J Clin Oncol* 1983; 1: 710–19.
10. Buyse M, On the relationship between response to treatment and survival time. *Stat Med* 1996; 15: 2797–812.
11. Murray N, Coppin C, Coldman A et al, Drug delivery analysis of the Canadian multicenter trial in non-small cell lung cancer. *J Clin Oncol* 1994; 12: 2333–9.
12. Seidman AD, Portenoy R, Yao T-J et al, Quality of life in phase II trials: a study of methodology and predictive value in patients with advanced breast cancer treated with paclitaxel plus granulocyte colony-stimulating factor. *J Natl Cancer Inst* 1995; 87: 1316–22.
13. Coates AS, Gebski V, Bishop J et al, for the Australia–New Zealand Breast Cancer Trials Group, Improving the quality of life during chemotherapy for advanced breast cancer: a comparison of intermittent and continuous treatment strategies. *N Engl J Med* 1987; 317: 1490–5.
14. Sullivan BA, McKinnis R, Laufman LR, Quality of life in patients with metastatic colorectal cancer receiving chemotherapy: a randomized double blind trial comparing 5-FU versus 5-FU with leucovorin. *Pharmacotherapy* 1995; 15: 600–7.
15. Spilker B (ed), *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd edn. Lippincott-Raven: Philadelphia, 1996.
16. Shaughnessy JA, Wittes RE, Burke G et al, Commentary concerning demonstration of safety and efficacy of investigational anticancer agents in clinical trials. *J Clin Oncol*, 1991; 9: 2225–32.
17. Gelber RD, Bonetti M, Cole BF et al, Quality of life assessment in the adjuvant setting: is it relevant: In: *Adjuvant Therapy for Breast Cancer VI* (Senn H-J, Goldhirsch A, Gelber RD, Thurlimann B, eds). Springer-Verlag: Berlin, 1998: 373–89.

18. Aaronson NK, Ahmedzai S, Bergman B et al, The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993; 84: 365–76.
19. Cella DF, Tulsky DS, Gray et al, The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *J Clin Oncol* 1993; 11: 570–9.
20. Bernhard J, Hürny C, Coates AS et al, for the International Breast Cancer Study Group (IBCSG), Quality of life assessment in patients receiving adjuvant therapy for breast cancer: the IBCSG approach. *Ann Oncol* 1997; 8: 825–35.
21. Spitzer WO, Dobson AJ, Hall J et al, Measuring the quality of life of cancer patients. *J Chron Dis* 1981; 34: 585–97.
22. Torrance GW, Furlong W, Feeny D et al, Multi-attribute preference functions: Health Utilities Index. *Pharmacoeconomics* 1995; 7: 503–20.
23. Weeks J, O'Leary J, Fairclough D et al, The 'Q-tility' index: a new tool for assessing health-related quality of life and utilities in clinical trials and clinical practice. *Proc Am Soc Clin Oncol* 1994; 13: 436.
24. Bernhard J, Gelber RD (Guest Eds), Workshop on Missing Data in Quality of Life Research In cancer Clinical Trials: Practical and Methodological Issues. *Stat Med* 1998; 17: 511–793.
25. Goldhirsch A, Gelber RD, Simes RJ et al, for the Ludwig Breast Cancer Study Group, Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *J Clin Oncol* 1989; 7: 36–44.
26. Gelber RD, Cole BF, Gelber S, Goldhirsch A, The Q-TWiST method. In: *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd edn (Spiker B, ed). Lippincott-Raven: Philadelphia, 1996: 437–44.
27. Fairclough DL, Fetting JH, Cella D et al, Quality of life and quality adjusted survival for breast cancer patients receiving adjuvant therapy. *Quality of Life Res* 1999; in press
28. Gelber RD, Cole BF, Goldhirsch A et al, Adjuvant chemotherapy plus tamoxifen compared with tamoxifen alone for postmenopausal breast cancer: a meta-analysis of quality-adjusted survival. *Lancet* 1996; 347: 1066–71.
29. Cole BF, Gelber RD, Kirkwood JM et al, Quality-of-life-adjusted survival analysis of interferon alfa-2b adjuvant treatment of high-risk resected cutaneous melanoma: an Eastern Cooperative Oncology Group study. *J Clin Oncol* 1996; 14: 2666–73.
30. Gelber RD, Goldhirsch A, Cole BF et al, A quality-adjusted time without symptoms or toxicity (Q-TWiST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. *J Natl Cancer Inst* 1996; 88: 1039–45.
31. Brown ML, Nayfield SG, Shibley LM, Adjuvant therapy for stage III colon cancer: economics returns to research and cost-effectiveness of treatment. *J Natl Cancer Inst* 1994; 86: 424–30.
32. Neymark N, Kiebert W, Torfs K et al, Methodological and statistical issues of quality of life (QOL) and economic evaluation in cancer clinical trials: report of a workshop. *Eur J Cancer* 1998; 34: 1317–33.
33. Gold M, Russel L, Siegel J, Weinstein M, *Cost-Effectiveness in Health and Medicine*. Oxford University Press: New York, 1996.
34. Falkson CI, Ibrahim J, Kirkwood JM et al, Phase III trial of dacarbazine versus dacarbazine with interferon α-2b versus dacarbazine with tamoxifen versus dacarbazine with interferon α-2b and tamoxifen in patients with metastatic malignant melanoma: an Eastern Cooperative Oncology Group study. *J R Clin Oncol* 1998; 16: 1743–51.
35. Kaplan EL, Meier P, Nonparametric estimation from incomplete observations. *J Am Statist Assoc* 1958; 53: 457–81.
36. Cox DR, *The Analysis of Binary Data*. Methuen Press: London, 1972.
37. Cox DR, Regression models and life-tables (with discussion). *J R Statist Soc B* 1972; 34: 187–220.

38. ASCO Special Article, Critical role of phase I clinical trials in cancer treatment. *J Clin Oncol* 1997; 15: 853–9.
39. Ratain MJ, Mick R, Schilsky RL, Siegler M, Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. *J Natl Cancer Inst* 1993; 83: 1637–43.
40. Simon R, Freidlin B, Rubinstein L et al, Accelerated titration designs for phase I clinical trials in oncology. *J Natl Cancer Inst* 1997; 89: 1138–47.
41. Von Hoff DD, Turner J, Response rates, duration of response and dose response effects in phase I studies of antineoplastics. *Invest New Drugs* 1991; 9: 115–22.
42. Arbuck SG, Workshop On phase I design. Ninth NCI–EORTC New Drug Development Symposium, Amsterdam, March 12, 1996. *Ann Oncol* 1996; 7: 567–73.
43. Penta JS, Rosner GL, Trump DL, Choice of starting dose and escalation for phase I studies of antitumor agents. *Cancer Chemother Pharmacol* 1992; 31: 247–50.
44. Siu LL, Rowinsky EK, Clark GM et al, Dose escalation using the modified continuous reassessment method (MCRM) in phase I clinical trials: a review of the San Antonio experience. *Ann Oncol* 1998; 9(Suppl 2): 127 (abst).
45. Lebecq A, Dieras V, Marty M et al, Preliminary evaluation of the modified continuous reassessment method (MCRM) in four phase I studies with RPR 109881A. *Ann Oncol* 1998; 9(Suppl 2): 127 (abst).
46. Collins JM, Zaharko DS, Dedrick RL, Chabner BA, Potential roles for preclinical pharmacology in phase I trials. *Cancer Treat Rep* 1986; 70: 73–80.
47. Collins JM, Grieshaber BA, Pharmacologically guided phase I trials based upon preclinical development. *J Natl Cancer Inst* 1990; 82: 1321–6.
48. Graham MA, Kaye SB, New approaches in preclinical and clinical pharmacokinetics. In: *Cancer Surveys*, Vol 17: *Pharmacokinetics and Cancer Chemotherapy*. Imperial Cancer Research Fund: London, 1993: 27–49.
49. O'Quigley J, Pepe M, Fisher L, Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990; 46: 33–48.
50. O'Quigley J, Estimating the probability of toxicity at the recommended dose following a phase I clinical trial in cancer. *Biometrics* 1992; 48: 853–62.
51. Goodman SN, Zahurak ML, Piantadosi S, Some practical improvements in the continual reassessment method for phase I studies. *Stat Med* 1995; 14: 1149–61.
52. Mick R, Ratain MJ, Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. *J Natl Cancer Inst* 1993; 85: 217–23.
53. Babb J, Rogatko A, Zacks S, Cancer phase I trials: efficient dose escalation with overdose control. *Stat Med* 1998; 17: 1103–20.
54. Dent SF, Eisenhauer EA, Phase I trial design: Are new methodologies being put into practice? *Ann Oncol* 1996; 7: 561–6.
55. Eisenhauer EA, Phase I and II trials of novel anticancer agents: endpoints, efficacy and existentialism. The Michel Clavel Lecture. *Ann Oncol* 1998; 9: 1047–52.
56. Boral AL, Dessain S, Chabner BA, Clinical evaluation of biologically targeted drugs: obstacles and opportunities. *Cancer Chemother Pharmcol* 1998; 42(Suppl): S3–21.
57. Ettinger DS, Evaluation of new drugs in untreated patients with small-cell lung cancer: its time has come. *J Clin Oncol* 1990; 8: 374–7.
58. Wells RJ, Phase II window therapy. *J Clin Oncol* 1995; 13: 302.
59. Ettinger DS, Finkelstein DM, Abeloff MD et al, Justification for evaluating new anticancer drugs in selected untreated patients with extensive-stage small cell lung cancer: an Eastern Cooperative Oncology Group randomized study. *J Natl Cancer Inst* 1992; 84: 1077–84.
60. Gehan EA, The determination of the number of patients in a preliminary and a

follow-up trial of a new chemotherapeutic agent. *J Chron Dis* 1961; 13: 346–53.
61. Fleming TR, One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; 38: 143–51.
62. Simon R, Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials* 1989; 10: 1–10.
63. Simon R, How large should a phase II trial of a new drug be? *Cancer Treat Rep* 1987; 71: 1079–85.
64. Simon R, Wittes RE, Ellenberg SS, Randomized phase II clinical trials. *Cancer Treat Rep* 1985; 69: 1375–85.
65. Weitz JJ, Jung SH, Marschke RF Jr et al, Randomized phase II trial of two schedules of topotecan for the treatment of advanced stage non-small cell lung carcinoma (NSCLC): a North Central Cancer Treatment Group (NCCTG) trial. *Proc Am Soc Clin Oncol* 1995; 14: A1053.
66. Hoskins P, Eisenhauer E, Beare S et al, A randomized phase II study of two schedules of topotecan in previously treated patients with ovarian cancer. *J Clin Oncol* 1998; 16: 2233–7.
67. Thall PF, Simon R, Estey EH, New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clin Oncol* 1996; 14: 296–303.
68. Zee B, Melnychuk D, Dancey J et al, A new design of phase II cancer clinical trials incorporating response and early progression. *Controlled Clin Trials* 1996; 17: 85S (abst).
69. Dent S, Zee B, Dancey J et al, Design of phase II clinical trials – stopping rule using response & early progression. *Ann Oncol* 1996; 7(S1): 134 (abst).
70. Gwyther SJ, Gore ME, ten Bokkel Huinink W et al, Results of independent radiological review of over 400 patients with advanced ovarian cancer treated with topotecan (Hycamtin). *Proc Am Soc Clin Oncol* 1997; 16: 351a.
71. Flaherty LE, Liu PY, Unger J, Sondak VK, Comparison of patient characteristics and outcome between a single-institution phase II trial and a cooperative-group phase II trial with identical eligibility in metastatic melanoma. *Am J Clin Oncol* 1997; 20: 600–4.
72. Rasmussen H, Rugg T, Brown P et al, A 371 patient meta-analysis of studies of marimastat in patients with advanced cancer. *Proc Am Soc Clin Oncol* 1997; 16: 429a.
73. Zelen M, The randomization and stratification of patients to clinical trials. *J Chron Dis* 1974; 27: 365–75.
74. Pocock SJ, Simon R, Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; 31: 103–15.
75. Pocock S, *Clinical Trials: A Practical Approach*. Wiley: New York, 1983.
76. Peterson B, George SL, Sample size requirements and length of study for testing interaction in a $2 \times k$ factorial design when time-to-failure is the outcome. *Controlled Clin Trials* 1993; 14: 511–22.
77. International Breast Cancer Study Group, Effectiveness of adjuvant chemotherapy in combination with tamoxifen for node-positive postmenopausal breast cancer patients. *J Clin Oncol* 1997; 15: 1385–94.
78. Brown BW, The crossover experiment for clinical trials. *Biometrics* 1980; 36: 69–79.
79. Blackwelder WC, 'Proving the null hypothesis' in clinical trials. *Controlled Clin Trials* 1982; 3: 345–53.
80. Fleiss JL, *Statistical Methods for Rates and Proportions*. 2nd edn. Wiley: New York, 1981.
81. George SL, Desu MM, Planning the size and duration of a clinical trial studying time to some critical event. *J Chron Dis* 1974; 27: 15–24.
82. Freedman LS, Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982; 1: 121–9.
83. O'Brien PC, Fleming TR, A multiple testing procedure for clinical trials. *Biometrics* 1979; 35: 549–56.
84. Lan KKG, DeMets DL, Discrete sequential

boundaries for clinical trials. *Biometrika* 1983; 70: 659–63.
85. Pampallona S, Tsiatis AA, Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J Statist Planning Infer* 1994; 42: 19–35.
86. *EaSt: A Software Package for the Design and Interim Monitoring of Group Sequential Clinical Trials*. Cytel Corporation: Cambridge, MA, 1993.
87. Smith MA, Ungerleider RS, Korn EL et al, Role of independent data-monitoring committees in randomized clinical trials sponsored by the National Cancer Institute. *J Clin Oncol* 1997; **15**: 2736–43.
88. Zelen M, Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. *J Clin Oncol* 1983; **1**: 164–9.
89. Begg C, Cho M, Eastwood S et al, Improving the quality of reporting of randomized controlled trials: the CONSORT (consolidation of standards for reporting trials) statement. *JAMA* 1996; 276: 637–9.
90. Peto R, Statistical aspects of cancer trials. In: *Treatment of Cancer* (Halnan KE, ed). Chapman and Hall: London, 1982: 867–71.
91. Glass GV, Primary, secondary and meta-analysis of research. *Educational Researcher* 1976; **5**: 3–8.
92. Begg CB, Berlin JA, Publication bias: a problem in interpreting medical data. *J R Statist Soc A* 1988; **151**: 419–63.
93. Early Breast Cancer Trialists' Collaborative Group, Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. *N Engl J Med* 1988; **319**: 1681–92.
94. Early Breast Cancer Trialists' Collaborative Group, *Treatment of Early Breast Cancer*, Vol 1: *Worldwide Evidence 1985–1990*. Oxford University Press: Oxford, 1990.
95. Early Breast Cancer Trialists' Collaborative Group, Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998; **351**: 1451–67.
96. Gelber RD, Goldhirsch A, Coates AS, for the International Breast Cancer Study Group, Adjuvant therapy for breast cancer: understanding the overview. *J Clin Oncol* 1993; **11**: 580–5.
97. Taylor KM, Margolese RG, Soskolne CL, Physicians' reasons for not entering patients in a randomized clinical trial of surgery for breast cancer. *N Engl J Med* 1984; **310**: 1363–7.
98. Llewellyn-Thomas HA, McGreal MJ, Thiel EC et al, Patients' willingness to enter clinical trials: measuring the association with perceived benefit and preference for decision participation. *Soc Sci Med* 1991; 32: 35–42.
99. Winn RJ, Obstacles to the accrual of patients to clinical trials in the community setting. *Semin Oncol* 1994; **21**: 112–17.
100. Gelber RD, Goldhirsch A, Can a clinical trial be the treatment of choice for patients with cancer? *J Natl Cancer Inst* 1988; 80: 886–7.