Clinical Trial Design for Anticancer Therapies

J. Jack Lee

The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

INTRODUCTION

Over thousands of years of history, medicine has evolved from a purely observational, anecdotal accumulation of empirical experience to a methodical, evidence-based scientific discipline. Clinical trials, which are prospective studies to evaluate the effect of interventions in humans under prespecified conditions, have become a standard and integral part in the development of new cancer therapies. Since the landmark paper of Frei *et al.* (1958) was published almost 50 years ago, cancer clinical trials have not only fuelled the development of numerous anticancer therapies but have also provided a fertile ground for basic scientific research and for the development of new designs and statistical methods to be used in testing new agents. A properly planned and executed clinical trial remains the most definitive tool for evaluating the effect and applicability of new anticancer therapies (Pocock, 1983; Friedman *et al.*, 1998; Green *et al.*, 2003; Teicher and Andrews, 2004; Piantadosi, 2005; Crowley and Ankerst, 2006).

BASIC ELEMENTS OF CLINICAL TRIAL DESIGN

A clinical trial is a study conducted prospectively, implementing a well-defined plan that is established before the study begins. This plan, the study protocol, should specify all the key components of the trial. The study protocol needs to include scientific rationales for the trial and the study objective and design, as well as specific steps to guide the conduct of the trial. General overviews for clinical trial design have been written by Gehan (1997), Lee *et al.* (2001), and Nottage and Siu (2002). The study protocol serves as an operation manual for conducting the trial by specifying the standard operating procedures. It is also a contract between the study participants, study investigators, scientific community, funding agencies, and the regulatory agencies, such as the Food and Drug Agency (FDA) and the National Cancer Institute (NCI) in the United States and the European Medicines Agency (EMEA) in Europe. A protocol should include the nine basic elements of clinical trial design outlined below.

1. *Study objective:* The first and foremost requirement in the design of a clinical trial is for the investigator to declare the objective of the study. The objective could be to conduct a pilot/feasibility study of a new agent or procedure, to evaluate a drug's safety and define its toxicity profile, to determine a drug's efficacy on patient outcome or survival, or to determine an agent's impact on the patient's quality of life. Quantitatively, the study objective can be posed in a form of estimation or hypothesis testing. An example of an estimation problem is a study objective to quantify the clinical response rate for an inhibitor of the epidermal growth factor receptor (EGFR) in patients with untreated, non–small cell lung cancer (NSCLC). An example of a hypothesis testing problem is a study objective to determine the difference in clinical response rates between patients with breast cancer who are treated with a selective oestrogen receptor modulator (SERM) and those treated with an aromatase inhibitor (AI). In this example, the null hypothesis may be equal clinical response rates for patients in the two treatment arms. Depending on the purpose, clinical trials can also be classified as superiority trials, non-inferiority trials, equivalence trials, and so on.
2. *Study population:* The study population of a trial is typically defined by specifying the eligibility criteria, which list the inclusion criteria and the exclusion criteria. A more homogeneous study population allows a treatment to be evaluated in more comparable settings, but can restrict patient entry and the generalizability of the study conclusion. In the early phase of drug development, a heterogeneous population can be sought to assess a drug's toxicity, while a more homogeneous population will be

The Cancer Handbook 2nd *Edition*. Edited by Malcolm R. Alison

required in the later stages of drug development to assess a drug's efficacy in more specific settings.

3. *Type of study design:* The simplest design is a single-arm, open label study. Its result can be compared to a fixed, preset target or existing data from historical controls. A controlled clinical trial compares the effect of an intervention against a control group, which can vary in its definition. A concurrent control group can be one that receives no treatment, a placebo, or a different treatment regimen in the same trial (labelled as no treatment, placebo, or active control, respectively). The result from controlled clinical trials is more credible than that from trials using a single treatment arm without a control group or with only a historical control, because the parallel design in a controlled clinical trial can remove many potential confounders. A crossover trial uses the patients as their own controls, comparing their pretreatment states to their post-treatment states. Owing to the fact that cancer is a progressive disease and a cancer patient's status rarely returns to the same pretreatment condition after therapy, the crossover design is rarely used in oncology. Another useful design is the factorial design, which allows simultaneous evaluation of the main effect and the combination effect from multiple treatments. With either no interaction or a positive interaction in efficacy between two agents, a 2 × 2 factorial design is an efficient design compared to two single-arm studies. The benefit of a factorial design becomes more evident when the two agents have non-overlapping toxicity profiles. The factorial design is commonly used in cancer prevention settings for evaluating the preventive effects of multiple regimens and their combinations. For example, the effects of aspirin and β-carotene on cardiovascular and cancer prevention end points were studied in the Physicians' Health Study (Hennekens and Eberlein, 1985); the effects of α-tocopherol and β-carotene on lung cancer prevention were evaluated in a trial of the Alpha-Tocopherol, Beta-Carotene (ATBC) Cancer Prevention Study Group (Heinonen *et al.*, 1994); and the effects of β-carotene, vitamins E and C, and multivitamins were evaluated in the Physicians' Health Study II (Christen *et al.*, 2000).
4. *Randomization and blinding:* One of the biggest obstacles that investigators must overcome in order to reach a valid study conclusion is *bias*. Bias can creep into clinical trials in many different ways. For example, *selection bias* occurs in a trial when more patients who have good (or bad) prognoses are enrolled in the trial. This bias can produce better (or worse) results than expected, regardless of the attributes of the treatment. Evaluation bias occurs when an investigator knows a patient's treatment assignment, and the knowledge consciously or subconsciously affects the investigator's judgement of whether to deem the patient's outcome a response or not. In addition, a trial participant who knows that he/she is receiving placebo may have a greater likelihood of dropping out of the trial or that participant may choose to "drop in" to the treatment group by taking the active drug when it is available outside of the trial. In either situation, a valid conclusion would be impeded because of bias. Randomizing the assignment of patients to treatment groups as they are enrolled in a trial and blinding the investigators and patients to treatment assignments are two very effective methods to guard against bias in clinical trials. Randomization can produce comparable study groups with respect to the known or unknown risk factors. It removes the potential allocation bias by taking out the subjectivity in treatment assignment. Proper randomization also guarantees the validity of statistical inference, for example, by applying a permutation test. Whenever possible, randomization should be implemented for any trial with two or more treatment groups.

   Several randomization schemes are useful in clinical trials. Simple randomization allocates patients into treatment arms with fixed probabilities, for example, patients can be assigned to treatment A or treatment B with 1 : 1 or 2 : 1 randomization ratios. When strong prognostic factors are known, patients should be stratified into various prognostic groups first and then randomized into treatment arms. Stratified randomization increases the chance of balanced treatment assignment within each prognostic group. In addition, a random permuted block design can guarantee a balance in treatment allocation. For example, a randomization list using a random permuted block design with two treatments (A and B) and a block size of four may look like the following: ABBA, BABA, AABB, BAAB, and so on. It may not be possible to balance randomization within each subgroup when there are large numbers of prognostic factors. For example, with five prognostic factors and three levels for each factor, there are $3 \times 3 \times 3 \times 3 \times 3 = 243$ subgroups. If the target accrual of a trial is only 100, there will not be enough subjects in each subgroup to perform stratified randomization. In this case, minimization methods (minimizing a predefined imbalance score), such as the dynamic allocation method (Pocock and Simon, 1975), can be useful in achieving "marginal" balance for each of the five prognostic factors, but not for their combinations. These randomization methods can be used alone or together whenever applicable (Rosenberger and Lachin, 2002).

   Study blinding is also important for preserving the integrity of the trial. As mentioned previously, the study result can be biased when the investigator and/or the patient knows the treatment assignment. Therefore, single-blinded trials (treatment assignment is blinded to either patients or investigators) or double-blinded trials (treatment assignment is blinded to both patients and investigators) are designed to alleviate these problems. In cancer therapy, it may not be possible to "blind" the treatment to patients, physicians, or both because the routes of treatment administration for many anticancer therapies and/or their distinct toxicity profiles are obvious. In this case, at the very least, the person who evaluates the treatment outcome (e.g., the radiologist or the pathologist) should be blinded to the treatment assignment to avoid evaluation bias.
5. *Primary/secondary end point:* Every clinical trial should have at least one clearly stated primary end point that is

consistent with the primary objective of the trial. In the early phase of drug development (phase I), the highest effective dose with acceptable toxicity, the frequency, and the profile of toxicities could be the primary end points. In the middle phase of drug development (phase II), a short-term clinical end point, such as response rate, disease-free survival, or time to disease progression, is typically used as the primary end point. In the late phase of drug development (phase III), the overall survival is often used as a gold standard primary end point to confirm whether the new agent imparts a real benefit. On the basis of the primary end point, sample size can be determined to ensure that a definitive conclusion can be drawn at the completion of the study. Secondary end points should also be defined prospectively so that the data can be properly collected and analysed.

6. *Power and sample size calculation:* Under the hypothesis testing framework, the sample size required for a study can be calculated by specifying the type I error rate (α or false-positive rate), statistical power (or $1 - \beta$, where β is type II error or false-negative rate), target difference to be detected, magnitude of variability, and the statistical test to be used. Typically, the type I error (could be one sided or two sided) rate is set at 5 or 10%, and the statistical power is set at 80% or higher to claim a definitive study result. In the estimation setting, the sample size can be calculated by specifying a desirable precision for an estimator (e.g., response rate) such that the standard error is controlled or such that it yields a 95% confidence interval within a given length.
7. *Accrual plan:* Although cancer is a common disease, with a lifetime cancer risk of 1 : 2 in men and 1 : 3 in women (Jemal *et al.*, 2006). Cancer trials can have difficulty in enrolling patients because less than 10% of cancer patients participate in clinical trials. This low rate of participation has been attributed to a lack of suitable protocols, stringent eligibility criteria, and reluctance on the part of some physicians to engage in trial accrual (Tannock, 1995; Comis *et al.*, 2003). One of the most common reasons for the failure of a clinical trial is slow accrual. Therefore, the accrual plan needs to be well thought out before the trial begins. The projected accrual rate and study duration should be realistic and clearly stated in the study protocol. Factors affecting accrual, such as the therapeutic potential of the new agent(s) and the socio-demographic status of patients in the potential study population and their access to health care, must be carefully considered beforehand in order to meet the anticipated study accrual (Nurgat *et al.*, 2005; Baquet *et al.*, 2006).
8. *Interim monitoring plan:* Most cancer clinical trials take years to complete – even accruing sufficient numbers of patients may take several years. It is important to implement interim monitoring plans for trials that last 3 or more years. On the basis of the available interim results, timely analyses can help to make appropriate adjustments to the trial or to terminate the study early because of toxicity or convincing efficacy or futile (lack of efficacy) results. Any interim analyses should be planned before the trial starts and monitored by an independent body (such as a data monitoring committee (DMC) or a data and safety monitoring board (DSMB)), which can provide oversight of the conduct of the study and give recommendations. Interim analyses should be designed to preserve the type I and type II error rates. Properly designed interim analyses can gain efficiency for the study and fulfill ethical requirements of not subjecting patients to ineffective and/or toxic treatments once such interim results are known.
9. *Data analysis plan:* The plan for statistically analysing both the primary and secondary end points should be laid out in the study protocol. For randomized clinical trials, intent-to-treat (ITT) analysis refers to the analysis of patients according to their assigned treatment arm, regardless of whether the patient received the assigned treatment or not, or how much of the assigned treatment the patient received. Compared to analyses based on evaluable patients or only those patients who completed the assigned treatment, ITT analysis is more conservative, introduces less bias, and reflects the realistic treatment effect when the treatment is delivered to the target population.

PARADIGM FOR CANCER CLINICAL TRIAL DESIGN

The development of anticancer agents can be divided into two phases: the preclinical and the clinical phase. The preclinical phase includes the identification or synthesis of compounds, *in vitro* testing in cell lines, and *in vivo* testing in animals. Before moving to the first study involving humans, the investigators should have a good understanding of a drug's mechanism of action, pharmacokinetics (PK), pharmacodynamics (PD), and toxicity profile based on the *in vitro* and *in vivo* studies. Animal studies are essential to identify the LD_{10} (the dose that kills 10% of the experimental animals) in mice and/or the TD_{Lo} (the toxic dose corresponding to the lowest dose that produces any toxic effect in the experimental animals) in dogs. After passing the preclinical phase, the drug enters the clinical phase of drug development, which can be classified as phase I, phase II, and phase III studies. Phase IV studies are devoted to post-marketing surveillance of the use of drugs in specific populations. The characteristics of clinical trials by phase are listed in Table 1, and the corresponding clinical trial designs are described in the following sections.

Phase I Designs: Clinical Pharmacology and Toxicity

To develop any new drugs or combinations and to determine the dose/schedule, safety must be established first. Phase I studies are employed to evaluate a drug's toxicity profile and to collect clinical pharmacology data to understand the PK and PD of new agents. For a cytotoxic agent, the primary goal of a phase I study is typically to identify the

Table 1 The characteristics of clinical trials by drug development phases.

	Phase I	Phase II	Phase III	Phase IV
Main purpose	Clinical pharmacology and toxicity	Initial assessment of efficacy	Full-scale evaluation of treatment efficacy	Post-marketing surveillance
End point	Dose-limiting toxicity	Short-term clinical response	Overall survival, quality of life	Comprehensive, long-term toxicity end points
	Maximum tolerated dose	Response rate (complete and partial response)		
	Pharmacokinetics	Disease-free survival		
	Pharmacodynamics	Time to progression		
Sample size	15–30	30–100	>100	>10 000
Patient characteristics	Sicker		Healthier	All subjects take the drugs
	Heavily pretreated		No prior treatment	
	Failed standard treatment		Good prognosis	
	Poor prognosis		Homogeneous	
	Heterogeneous			
Intended use	Second-, third-, or fourth-line therapy		Front-line therapy	

maximum tolerated dose (MTD) in a dose escalation fashion. Key elements of phase I studies include: (i) defining the starting dose, (ii) defining the toxicity profile and dose-limiting toxicity (DLT), (iii) defining an acceptable level of toxicity or the target toxicity level (TTL), and (iv) defining the dose escalation scheme. (i) Defining the starting dose: For the first study in humans, the starting dose is usually chosen as one-tenth of the LD_{10} in mice or one-third of the TD_{Lo} in dogs, as these dose levels have been shown to be safe in humans. Using a safe starting dose is important; however, the investigator must balance the risk of toxicity with the risk of treating patients with ineffective doses, which would happen if the starting dose were too low to have a biologic effect. (ii) Defining the toxicity profile and dose-limiting toxicity (DLT): Some level of toxicity is almost always expected for anticancer agents, thus investigators of cancer therapy must include guidelines in their study protocols for preventing, measuring, and managing their patients' adverse reactions from drug toxicity. Although animal studies may provide clues, it is only when agents are introduced into clinical trials that possible toxicities in humans can be characterized. Most grade 1 or 2 toxicities are transient and tolerable, but toxicities of grade 3 or higher tend to be more serious and/or non-reversible (according to NCI CTCAE 3.0 (http://ctep.cancer.gov/forms/CTCAEv3.pdf) or WHO toxicity criteria (http://www.fda.gov/cder/cancer/toxicityframe.htm)). In general, grade 3 non-haematological toxicities and grade 4 haematological toxicities are considered dose limiting. Depending on specific agents and study populations, each trial is required to specify the exact definition of the DLTs. (iii) Defining an acceptable level of toxicity or the TTL: The commonly acceptable level of toxicity in clinical trials is between 20 and 33%. Although the acceptable toxicity level is implicitly defined in many trials, an explicit definition is desirable. (iv) Defining the dose escalation scheme: The dose escalation scheme contains three components: (a) dose spacing, (b) dose assignment, and (c) cohort size. After specifying the starting dose, the investigator needs to specify the subsequent dose levels, how patients will be assigned to the various doses, and the size of each patient cohort. Some studies use a predetermined, fixed spacing of dose levels, such as 10, 20, and then 30 mg, and so on. Other trials may only specify a general scheme, such as doubling the dose when no toxicities are observed, reducing to a 50% dose escalation when non–dose-limiting grade 2 or higher toxicities are observed, and reducing to a 25% dose escalation when a DLT is observed. A modified version of the Fibonacci scheme is commonly used to determine the dosage scheme for a trial. The modified Fibonacci scheme corresponds to a series of dose increments in amounts of 100, 67, 50, 40, 33, 33, ..., 33%. In addition, the maximum dose level to be delivered in the trial should be stated. Dose assignment, which refers to the guideline on how to assign dose levels to patients subsequently enrolled in the trial, and cohort size for treating patients should also be given. We will discuss the commonly used algorithm-based methods and model-based methods next.

Algorithm-based Design: the 3 + 3 *Design*

A prevailing method for conducting phase I cancer clinical trials over the past decades has been the conventional 3 + 3 design or its variations. The algorithm for the 3 + 3 design is shown in Table 2. The design is simple and easy to implement but has its limitations. Although no specific TTL is defined, the 3 + 3 design generally yields an MTD that corresponds to a dose with 20 to 30% DLTs in treated patients (Smith *et al.*, 1996; Lin and Shih, 2001). The algorithm needs to be modified if a different TTL is set. The algorithm-based method is inefficient because the decision of dose escalation or de-escalation depends only on the data at the current dose level and ignores information from all other dose levels. If the starting dose is much too low, the 3 + 3 design can take a long time to reach the MTD.

Pharmacologically Guided Dose Escalation Design

Similar to the 3 + 3 design, the pharmacologically guided dose escalation (PGDE) design (Collins *et al.*, 1990; Fuse

Table 2 The conventional, algorithm-based 3 + 3 design.

Step 1:	Enter 3 patients at the lowest dose level
Step 2:	Observe the toxicity outcome
	0/3 DLT → Treat next 3 patients at next higher dose
	1/3 DLT → Treat next 3 patients at the same dose
	1/3 + 0/3 DLT → Treat next 3 patients at next higher dose
	1/3 + 1/3 DLT → Define this dose as MTD or exceeding MTD
	1/3 + 2/3 or 3/3 DLT → Exceeding MTD
	2/3 or 3/3 DLT → Exceeding MTD
Step 3:	Repeat step 2 until MTD is reached. If the last dose exceeds MTD, define the previous dose level as MTD if 6 or more patients were treated at that level or if treating more patients at the previous or intermediate dose level
Step 4:	MTD is defined as a dose with $\leq$1/6 DLT or $\leq$2/6 DLT

et al., 1994) bases the starting dose selection on relevant preclinical data, but measures the PK data to determine the subsequent dose spacing. The PGDE design allows a dose escalation of up to 100%, until the area under the curve (AUC) is within 40% of the target AUC, and then it switches to the modified Fibonacci scheme to determine dose spacing.

Model-based Design: Continual Reassessment Method

The continual reassessment method (CRM) (O'Quigley *et al.*, 1990) is a model-based method used to estimate the dose–toxicity curve and, subsequently, to identify the MTD. A TTL (e.g., 20, 30, or 33% of patients developing DLT) and a one-parameter dose–toxicity curve need to be specified in advance. Let θ denote the probability of DLT; d, the dose; and a, the parameter of the dose–toxicity curve. Three families of dose–toxicity curves have been proposed:

Hyperbolic tangent model:

$$\theta = [\exp(d)/(\exp(d) + \exp(-d))]^a$$

Logistic model: $\theta = \exp(3 + a \cdot d)/(1 + \exp(3 + a \cdot d))$

Power model: $\theta = d^{\exp(a)}$

Figure 1 shows the dose–toxicity curves for these three models. After specifying a prior distribution of a, the dose closest to the TTL can be determined for treating the next patient. Upon observing the toxicity outcome, the posterior distribution of a can be calculated. As the trial moves along, the dose–toxicity curve is refined at each step by updating the posterior distribution of a. The CRM allows the investigator to treat patients at the dose closest to the current estimate of the MTD, thereby maximizing the chance for patients to receive putatively the most efficacious dose within the bounds of pre-specified toxicity on the basis of the current data. Subsequent modifications proposed for the CRM incorporate additional safety measures, such as starting with the lowest dose level, not skipping dose levels, and so on (Korn *et al.*, 1994; Goodman *et al.*, 1995). The advantages of the model-based methods include (i) a clearly defined objective of the TTL; (ii) a more rapid dose escalation, which allows patients to be treated at doses close to the target MTD level, such that the number of patients treated at low or ineffective dose levels can be reduced; and (iii) the use of all available information in determining the MTD. The drawback is that these methods are more complicated and require special software for their implementation. An excellent tutorial of the CRM can be found in Garrett-Mayer (2006).

Accelerated Titration Design

The accelerated titration design (ATD) (Simon *et al.*, 1997) starts by treating one patient at the lowest dose level. In the accelerated titration phase, the method allows 100% inter- and intrapatient dose escalation when no toxicities or only grade 1 toxicities are observed. The standard design phase kicks in after observing grade 2 toxicities during any course (cycle). The standard design phase uses the 3 + 3 design with 40% dose escalation. The design allows intrapatient dose escalation in multiple courses of the same patient in the following ways. (i) When no toxicity or only grade 1 toxicity is observed, the patient is treated with a 100% dose increase in the next course. (ii) When grade 2 toxicity is observed, the patient is treated at the same dose level in the next course. (iii) When grade 3 or 4 toxicity is observed, the patient is treated with a dose level that is decreased by one dose level in the next course. All toxicity information is then used to model the dose–toxicity curve.

Several other methods have been proposed in the literature for phase I studies (see reviews by Storer, 1989; Ahn, 1998; Parulekar and Eisenhauer, 2004; O'Quigley and Zohar, 2006). The method of escalation with overdose control (EWOC) (Babb *et al.*, 1998; Rogatko *et al.*, 2005) allows for rapid dose escalation within the constraint of an estimated toxicity limit. The EffTox method allows for the specification of an efficacy–toxicity trade-off boundary and searches for the optimal dose within the specified boundaries using a Bayesian methodology (Thall and Cook, 2004).

Phase II Designs: Initial Assessment of Efficacy

After the toxicity profile and/or MTD for an anticancer treatment is determined, phase II studies are conducted to evaluate whether the new agent has sufficient anticancer activity and to refine the knowledge of its toxicity profile. For cytotoxic agents, phase II studies are often conducted at the MTD level or at the recommended phase II dose, which can be one dose lower than the MTD. For biological agents, phase II studies are often conducted at the optimal biological dose. Chronologically, a phase IIA trial is a single-arm study conducted to provide an initial efficacy assessment of a new agent with a goal of screening out ineffective drugs. The primary end point of a phase IIA trial is often the clinical response, defined as complete response (no evidence of disease) or partial response. Partial response is defined as tumour volume shrinkage of 50% or more based on a two-dimensional measurement or the one-dimensional response evaluation criteria in solid tumours (RECIST) (Therasse *et al.*, 2000, 2006). If an agent passes a phase IIA study, the

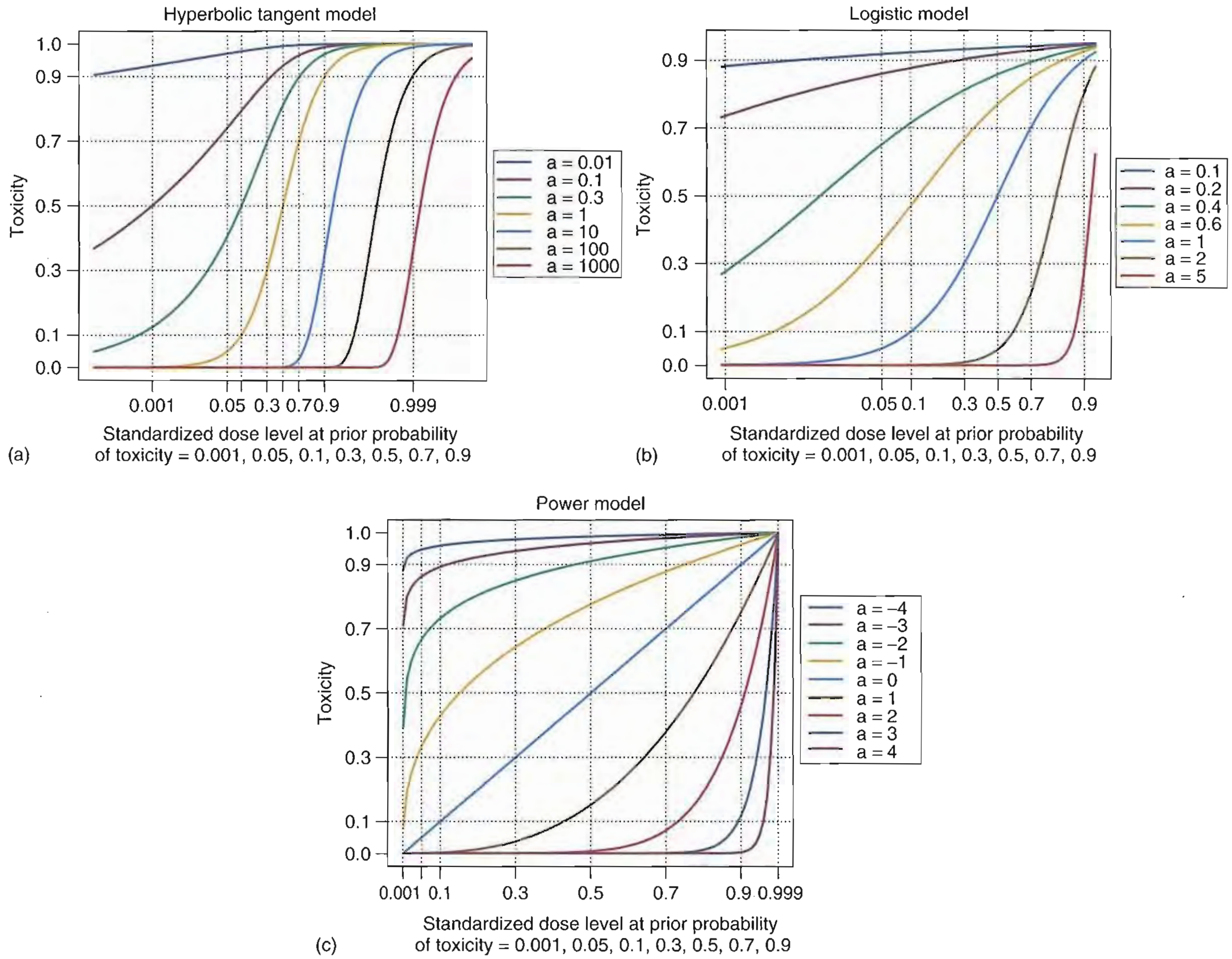

Figure 1 Family of dose–toxicity curves for the model-based continual reassessment method: (a) hyperbolic tangent model, (b) logistic model, and (c) power model.

subsequent phase IIB trial is often a randomized, multi-arm study with a goal of identifying the most promising treatment regimen to evaluate further in a phase III trial. Time to recurrence or time to progression can be used as the primary end point for a phase IIB trial. Comprehensive overviews on the design and analysis of phase II cancer trials can be found, for example, in papers by Mariani and Marubini (1996), Scher and Heller (2002), and Gray *et al.* (2006).

Phase IIA Designs

Typically, a phase IIA trial is a single-arm, open label study that requires the treatment of 30 to 100 patients under a multistage design. Some commonly used designs are described below.

Gehan's Design In the early days of anticancer drug development, there were few agents with any anticancer activity. A drug is considered active if it produces at least a 20% clinical response rate (p). To test the hypothesis of H_0: $p = 0$ versus H_1: $p = 0.2$, Gehan (1961) proposed a two-stage design. The first stage enrolls 14 patients. If none of them respond to the treatment, the drug is declared ineffective and the trial is stopped. If at least one clinical response is seen, additional patients (typically 20–40) are enrolled in the second stage to estimate the response rate with a prespecified precision. The design has a type I error rate of zero (because under the null hypothesis of $p = 0$, no response can occur) and 95% power for $p = 0.2$. The design can also be used to test other response rates under H_1. For example, for $p = 0.1$ or 0.15, the corresponding sample size in the first stage to achieve 95% power will be 29 or 19, respectively.

Simon's Two-stage Designs Simon (1989) proposed 2 two-stage designs, the optimal design and the minimal design, to test the hypotheses, H_0: $p \le p_0$ versus H_1: $p \ge p_1$, with given type I and type II error rates: These designs can be constructed to minimize the expected sample size or the maximum sample size under the null hypothesis, respectively. For example, when $p_0 = 0.1$ and $p_1 = 0.3$, with $\alpha = \beta = 0.1$, the optimal two-stage design enrolls 12 patients in the first stage. If no response or only one response is found, the trial is stopped and the agent is considered ineffective. Otherwise, 23 more patients are enrolled to reach a total of 35 patients. At the end of the trial, if only five or fewer responses are observed, the agent is deemed ineffective. Otherwise (i.e., with 6 or more responses in 35 patients), the agent is considered effective. Under the null hypothesis, there is a 66% chance that the trial will be stopped early. The expected sample size under H_0 is 19.8. In comparison, the minimax design enrolls 16 patients in the first stage. If no response or only one response is seen, the trial is stopped early and the agent is considered ineffective. Otherwise, 9 more patients are enrolled in the second stage to reach a total of 25 patients. At the end of the trial, the agent is considered ineffective if four or fewer responses are seen and effective otherwise. The expected sample size is 20.4 and the probability of early stopping is 0.51 under the null hypothesis. In both designs, the trial can be stopped early because of lack of efficacy (i.e., futility), to save patients from receiving ineffective treatments and to preserve other resources. If the treatment works well, there is little reason to stop the trial. More patients can be beneficially treated while the design continues to increase the precision in estimating the response rate. As it is unlikely that new agents always work as expected, a multi-stage design with rules for early stopping due to futility is desirable in phase II settings.

Other Multi-stage Designs Other multi-stage designs can be found in the literature. Fleming (1982) proposed a two-stage design that allows for early stopping due to futility or efficacy. Bryant and Day (1995) developed a two-stage design that allows the investigator to monitor efficacy and toxicity simultaneously. Three-stage designs were proposed by Ensign *et al.* (1994) and Chen (1997). Three-stage designs improve the efficiency of two-stage designs but are more complicated to implement and can increase the cost and length of the study. The gain in efficiency of designs with more than three stages does not justify the additional complexity in conducting such studies.

Phase IIB Designs

After an agent is shown to have certain anticancer activity in a phase IIA trial, its efficacy is compared with those of other active anticancer therapies in multi-arm, randomized phase II trials. Such studies can involve the same drug in more than one treatment arm when it is given in different doses or schedules or in combination with other drugs. The challenge is to choose the most promising regimens among a large number of potentially active regimens for further development. Lee and Feng (2005) reviewed 266 randomized phase II studies conducted from 1986 to 2002. They found that most studies applied randomization to achieve patient comparability, while embedding a one-sample phase II design within each treatment arm. Owing to limited sample sizes, such designs typically do not yield sufficient statistical power for a strict head-to-head comparison between treatment arms, as is possible in phase III trials. Notwithstanding this limitation, two other types of phase IIB designs, listed below, are available to allow some limited comparison of agents or regimens.

Pick-the-Winner Design The Simon, Wittes, and Ellenberg (SWE) method is based on the statistical methodology of ranking and selection with binary end points (Simon *et al.*, 1985). Unlike the ordinary hypothesis testing framework that controls both type I and type II errors, the ranking and selection procedure of the SWE method controls only type II errors. Basically, the response rate of each treatment arm is estimated and the arm with the highest response rate is picked as the winner and is sent forth for further evaluation. The design is appealing because the required sample size is much smaller than that for a randomized trial under the hypothesis testing framework. For example, $N = 146$ patients per arm are required for testing the response rates of 10 versus 25% with 90% power and a two-sided 5% type I error rate. On the other hand, the SWE method requires only $N = 21$ patients per arm with the same power. The trade-off, however, is that the false-positive rate can range from 20 to over 40%, as reported in simulation studies (Liu *et al.*, 1999). The SWE method works best when there is only one true "winner" with all other contenders falling below par. When there are several comparable, active regimens, the SWE method cannot accurately differentiate the best one from the better ones. At the end of the trial, this method always picks the treatment arm with the best observed outcome as the winner, regardless of whether none of the regimens work, some of them work, or all of them work well. In addition, another drawback of the SWE method is that it does not provide an early stopping rule due to futility. Therefore, there is no provision for terminating a non-performing arm early on the basis of interim results. Although the SWE method offers small sample sizes, the ranking and selection procedure does not fit well with the objectives for phase IIB studies, hence, only about 11% of the randomized phase II designs have used this method (Lee and Feng, 2005).

Comparative Designs Despite the enthusiasm of investigators, it is prohibitive to send every new agent for phase III evaluation due to the large number of active agents, limited number of patients, high cost, and great time commitment. To screen for the most active treatment in an intermediate step, randomized phase IIB studies are often designed with a moderate sample size (e.g., 50–200) and a medium study duration (e.g., 1–3 years). To offset the stringent type I and type II requirements that result in large sample sizes in phase III trials, the type I error rate is usually increased from 5 to 10 or even to 20%, while the maximum type II error rate is still controlled within 10 to 20%. The rationale is that in phase II trials, it is more important to control the type II error rate, or the false-negative rate, such that promising treatments will not be missed. A false-positive result is of less concern because the final verdict of the effectiveness of a regimen can be provided in a phase III evaluation. In addition, in order to shorten the study duration, earlier end points such as time to recurrence, disease-free survival, or time to progression are commonly employed in phase II settings. A moderate to large expected difference is often assumed for phase II studies.

Phase III Designs: Full-scale Evaluation of Treatment Efficacy

Phase III trials are considered definitive trials for comparing a new treatment with a standard treatment in a rigorous manner, for example, a double-blinded, randomized, placebo-controlled study. The goal is to define the best treatment, which implies a possible change in the current standard practice. Typically, stringent statistical requirements such as a two-sided 5% type I error rate and at least 80% power are required. The primary end point is often the overall survival rate. Reviews of sample size calculation can be found in papers by Julious *et al.* (1999), Julious (2004), and Julious and Patterson (2004). A phase III study requires hundreds of patients (in cancer treatment trials) or even thousands or tens of thousands of patients (in primary prevention trials). Generally speaking, a phase III trial tends to be a multicentre study with several years of accrual and follow-up and is therefore quite costly. Owing to the size and study duration, formal interim analyses, such as the group sequential methods by Pocock, O'Brien–Flemming, Peto, or Lan–Demets are often employed to assess early stopping needs due to efficacy, futility, or both (Geller, 1987; Geller and Pocock, 1987). A phase III trial is usually monitored by an independent, external DMC or DSMB to ensure the safety of study patients and to make recommendations based on the interim data. Compared to phase I or phase II studies, where only the evaluable patients are included in the analysis, most phase III studies employ the ITT principle in the data analyses. For new drug development, a phase III trial can be used as a registration trial to gather safety and efficacy data as a basis for drug approval.

Phase IV Designs: Post-marketing Surveillance

Following the approval and marketing of an anticancer drug, there is still a need to monitor the drug's adverse effects and the long-term morbidity and mortality of a large number of patients who were treated with the drug. Phase IV studies are thus required to define the long-term effects of a drug. The recent controversy surrounding the discovery of cardiovascular toxicities in patients treated with COX2 inhibitors has further underlined the importance of phase IV studies (Stern, 2003).

NOVEL CLINICAL TRIAL DESIGNS

Advancements in our understanding of cancer development and treatment on molecular and genetic levels have necessitated the proposal of many novel clinical trial designs in recent years. The general goal is to identify effective treatments and specific patient groups who may benefit from certain treatments in the most efficient way – to take a step towards personalized medicine (Jain, 2005; Ginsburg and Angrist, 2006).

Designs for Developing Targeted Agents

With a better understanding of cancer aetiology and the mechanism of action of anticancer treatments, target-based drug development has become more rational than the traditional methods based on empirical evidence and random screening. Instead of using toxicity and clinical response, as in the traditional phase I and II designs, it may be more suitable to use PK and biological end points, such as biomarker modulation or molecular imaging, to define the optimal biological dose in trials involving targeted agents. Nevertheless, measurable clinical benefit continues to be an important criterion for phase III trials (Korn *et al.*, 2001; Fox *et al.*, 2002; Parulekar and Eisenhauer, 2004; Schiller, 2004). Another important premise of targeted-agent development is that only a fraction of the patients may benefit from the targeted treatment. Hence, there is a need to enrich the study population such that treatments can be given to those who are more likely to respond, in order to gain efficiency in testing the efficacy of the targeted agents. Furthermore, a trial in which patients are treated with the agents most likely to work for them is a more ethical trial (Temple, 2005). An excellent review of clinical trial design and end point selection for the development of EGFR-targeted therapies can be found in a paper by Arteaga and Baselga (2003).

Efficient Targeted Randomized Designs

Many of the targeted agents are developed to "target" certain molecular defects. Presumably, patients presenting with the target will respond better to the targeted treatment, while the drug may not work at all or not work as well in patients without the target. Simon and Maitournam (2004) and Maitournam and Simon (2005) proposed efficient targeted randomized designs (ETRD) for targeted-agent development. When an accurate assay is available to screen for the target, it is generally better to screen for patients with the target and randomize only those patients. For example, investigators have postulated that gefitinib, recently used to treat non–small cell lung cancer, may work only in patients with an EGFR mutation (Lynch *et al.*, 2004; Paez *et al.*, 2004; Chou *et al.*, 2005). The presence of the mutation, identified through molecular analysis of a tissue sample, is then the target. Patients with that target would realize the most benefit from therapy with gefitinib and should be randomized to a trial for additional evaluation of this agent. Table 3 gives the required sample size in hypothetical situations. For example, when only 10% of the patients present with the target and the treatment has no effect in patients without the target ($\delta_0 = 0$), assuming the response rate is 40% in the standard treatment arm and 60% in

Table 3 Comparison between the general, untargeted design and the targeted design showing the sample size and efficiency (in parentheses).

Design	$\delta_0 = 0$		$\delta_0 = \delta_1/2$	
	$\delta_1 = 0.2$	$\delta_1 = 0.4$	$\delta_1 = 0.2$	$\delta_1 = 0.4$
Untargeted design	12 806	3248	446	116
Targeted design	138 (92.2)	34 (95.8)	138 (3.2)	34 (3.4)
Targeted design (screened)	1380 (9.2)	340 (9.6)	1380 (0.3)	340 (0.3)

the targeted treatment arm ($\delta_1 = 0.2$), 12 806 patients will be required to achieve 90% power with a two-sided 5% type I error rate. If we screen the patients and randomize only those patients who present with the target, we need to screen only 1380 patients to identify 138 patients to be randomized. The efficiency is 9.2 times for the patients who are screened and 92.2 times for the patients who are randomized. If the efficacy of the targeted treatment in patients without the target is half of that in patients with the target, then the untargeted design requires the enrollment of 446 patients. The same numbers of screened and randomized patients are required for the targeted design. The resulting efficiency is 0.3 for the number of patients screened and 3.2 for the number of patients randomized. These findings show that when a targeted treatment works only in a small fraction of patients, that is, those who present with the target, then the more efficient design is one that requires patient screening and enrolls only those patients who may benefit from the treatment. The results depend on the availability of a known assay with high sensitivity and specificity in order to accurately identify patients who will benefit from the treatment.

Randomized Discontinuation Design

A randomized discontinuation design (RDD) has been proposed to evaluate cytostatic agents to which only a fraction of patients respond to treatment and for which no predictive markers or assays are available to identify who will and will not respond (Rosner *et al.*, 2002). Similar to that of the ETRD, the objective of this targeted design is to enrich the study population. The proposed design can be carried out in two stages. In stage 1, all patients are given the treatment over a fixed period of time, for example, 4 months, and their responses to the treatment are then evaluated. In stage 2, patients who responded to the treatment in stage 1 (achieving complete or partial response) continue to receive the treatment, while patients experiencing disease progression are removed from the study and can receive alternative treatments. Patients achieving a stable disease state are randomized to either continue the treatment or discontinue. The efficacy of the targeted treatment can then be assessed in this group of patients. Compared with the trial design requiring up front randomization without patient selection, the RDD can be more efficient in certain cases. The RDD is less efficient if the treatment has a fixed effect on the tumour growth rate or if its only benefit is to slow the tumour growth rate. Simulation studies have shown, however, that up front randomization is generally more efficient (Capra, 2004; Freidlin and Simon, 2005a, 2005b). Successful implementation of the RDD has been described by Stadler *et al.* (2005).

Designs with Biomarker and Genomic End Points

In this genomic era, translational studies that integrate basic science studies with clinical trials are no longer the exception and have become quite common. Many clinical trials incorporate the collection of biospecimens and the analysis of biomarkers in the study objectives. Biomarkers can be considered prognostic (associated with disease outcome regardless of a particular treatment) or predictive (associated with disease outcome for a specific treatment) and can be included in the analysis of primary, secondary, or surrogate end points (Park *et al.*, 2004). Biomarker analysis can be incorporated in both prospective and retrospective studies. To standardize the evaluation of prognostic markers in the clinical setting, reporting recommendations for tumour marker prognostic studies (REMARK) were outlined, which include important elements such as patient/sample selection criteria, assay methods, study design, methods and results of statistical analyses, and conclusions (McShane *et al.*, 2005). In addition to examining several biomarkers, some trials now include genomic end points as the primary or secondary end points; for example, quantifying expression in thousands of genes using microarrays. Some investigators then use this information to establish a molecular profile of the disease and disease classifications, to screen for prognostic/predictive markers, to measure the treatment effect by marker modulation, and so on. Allison *et al.* (2006) provided an overview of the five key components in microarray experiments: design, preprocessing, inference, classification, and validation. Simon (2005) gave a road map for developing and validating genomic classifiers which are useful for selecting treatments for individual patients. The literature also contains many recent papers on design considerations including sample size calculation for such studies to ensure the control of the overall type I error rate (such as the family-wise error rate, the false-discovery rate, etc.), allowing the true difference to be discovered with sufficient statistical power (Lee and Whitmore, 2002; Pusztai and Hess, 2004; Wang and Chen, 2004; Dobbin and Simon 2005; Hu *et al.*, 2005; Jung *et al.*, 2005; Michiels *et al.*, 2005; Page *et al.*, 2006). The use of pharmacogenomics in drug discovery has also spurred much discussion (Penny and Mchale, 2005).

Bayesian Designs versus Frequentist Designs

Although the work of Thomas Bayes, known as Bayes' theorem, had been published posthumously as early as 1764, the frequentist methodology, which was not well developed until the work of R. A. Fisher in the early 1900s, has dominated the field of clinical trials in the past half century. The model

Table 4 Comparison between frequentist method and Bayesian method.

Property	Frequentist method	Bayesian method
Model framework	Probability of data/θ	Probability of θ/data
Data	Random	Fixed
Parameter (θ)	Fixed	Random
Inference	Hypothesis testing	Posterior distribution Bayes factor
	Confidence interval	Credible interval
Key feature	Control type I and type II error rates	Compute the operating characteristics
Conform to the likelihood principle	No	Yes
Require specifying prior distribution	No	Yes
Subjectivity in inference making	Less	More
Computation	Less intensive	More intensive
Software availability	More	Less
Adaptivity to complex problem	Less	More
Incorporate external information	Hard	Easy
Study design	Rigid	Flexible

framework of both approaches is shown in Table 4. The frequentist method is based on probability of data/θ, where data is considered random and the parameter θ is fixed. On the other hand, the Bayesian model is based on computing probability of θ/data, where the parameter θ is random while the data is given and fixed. The central dogma of the frequentist approach is to set up a null hypothesis and an alternative hypothesis, and then quantify the evidence supporting the data conditioned on the hypothesis. The commonly used p value is the probability of observing data that is as extreme or more extreme than the observed data, given that the null hypothesis is true. The null hypothesis is rejected if the p value is small. Such a setting of hypothesis testing allows investigators to control the type I (false positive) and type II (false negative) error rates, but does not directly calculate the probability of whether the null or the alternative hypothesis is true. In contrast, the Bayesian framework directly calculates the evidence supporting the null or the alternative hypothesis by computing the posterior probability. The posterior probability of a parameter(s) is computed by taking the product of the prior probability of the parameter(s) and the likelihood of data, given the parameter(s). Posterior probability can be considered as the synthesis of information contained in the prior distribution of the parameter and the data. The Bayesian method conforms to the likelihood principle, which states that all information in a sample is contained in the likelihood. Bayes factors and credible intervals can be computed to perform hypothesis testing and confidence interval estimations that are similar to those in the frequentist approach. The Bayesian method requires the specification of the prior distribution of parameters and is considered more subjective than the frequentist method, but has an advantage in that it allows other information to be easily incorporated in the analysis. The prior distribution should be specified in advance and a sensitivity analysis should be applied for the Bayesian analysis to ease the concern of achieving different conclusions from different prior specifications. Simulation studies should be performed with carefully chosen design parameters to calibrate the design such that desirable operating characteristics can be achieved. Designs can be constructed to control both the frequentist and Bayesian error rates (Wang *et al.*, 2005). The Bayesian method is more computationally intensive and there is less software available for its implementation in clinical trials. The computational aspect, however, is quickly changing owing to much improvement in both computing power and the development of efficient algorithms, such as the Markov chain Monte Carlo (MCMC) method. The Bayesian method is more flexible and adaptive by nature, even when the conduct of a study deviates from the original design. A deviation from the original design in the conduct of a study causes the frequentist properties to fall apart, but the Bayesian properties remain unchanged. Examples of additional advantages of the Bayesian design are that the Bayesian design (i) easily incorporates internal and external trial information, allowing for more informed inference and better decision making; (ii) allows the analyst to borrow strength (information) across different disease subgroups or similar treatments by constructing hierarchical models; (iii) facilitates the development of innovative trials such as seamless phase II/III trials, and (iv) allows treating more patients with effective agents with outcome-based adaptive randomization designs, and so on (Inoue *et al.*, 2002; Thall *et al.*, 2003; Berry, 2005, 2006).

CLINICAL TRIAL CONDUCT, REPORTING, PRACTICAL CONSIDERATIONS, AND USEFUL LINKS

In an effort to standardize the design, conduct, and reporting of clinical trials by integrating inputs from government (regulatory authority), industry, and academia, the International Conference on Harmonisation has issued guidelines (ICH E9) on the statistical principles of clinical trials (see Table 5) (ICH, 1999; Phillips and Haudiquet, 2003). The ICH consortium encompasses three main regions: Europe, Japan, and the United States. The guideline is quite comprehensive and can serve as a template for protocol development, study conduct, data analysis, and final reporting (Lewis, 1999; Lewis *et al.*, 2001).

Regarding clinical trial conduct, database development and remote data capture can facilitate and standardize data collection in clinical trials, especially in multicentre trials. The US National Cancer Institute has developed the Cancer Data Standards Repository (caDSR). Through this effort, common data elements (CDEs) are defined with an available CDE browser. More information can be found at the caDSR project site (http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr), which also provides a caDSR administration tool (http://cadsradmin.nci.nih.gov), CDE browser (http://cdebrowser.nci.nih.gov/CDEBrowser/), CDE curation tool (http://cdecurate.nci.nih.gov/cdecurate/), and

Table 5 Outline of International Conference on Harmonisation Guidelines of Statistical Principles for Clinical Trials (ICH E9).

1. Introduction
    - 1.1 Background and Purpose
    - 1.2 Scope and Direction
2. Considerations for Overall Clinical Development
    - 2.1 Trial Context
        - 2.1.1 Development Plan
        - 2.1.2 Confirmatory Trial
        - 2.1.3 Exploratory Trial
    - 2.2 Scope of Trials
        - 2.2.1 Population
        - 2.2.2 Primary and Secondary Variables
        - 2.2.3 Composite Variables
        - 2.2.4 Global Assessment Variables
        - 2.2.5 Multiple Primary Variables
        - 2.2.6 Surrogate Variables
        - 2.2.7 Categorized Variables
    - 2.3 Design Techniques to Avoid Bias
        - 2.3.1 Blinding
        - 2.3.2 Randomization
3. Trial Design Considerations
    - 3.1 Design Configuration
        - 3.1.1 Parallel Group Design
        - 3.1.2 Cross-over Design
        - 3.1.3 Factorial Designs
    - 3.2 Multi-centre Trials
    - 3.3 Type of Comparison
        - 3.3.1 Trials to Show Superiority
        - 3.3.2 Trials to Show Equivalence or Non-inferiority
        - 3.3.3 Trials to Show Dose-response Relationship
    - 3.4 Group Sequential Designs
    - 3.5 Sample Size
    - 3.6 Data Capture and Processing
4. Trial Conduct Considerations
    - 4.1 Trial Monitoring and Interim Analysis
    - 4.2 Changes in Inclusion and Exclusion Criteria
    - 4.3 Accrual Rates
    - 4.4 Sample Size Adjustment
    - 4.5 Interim Analysis and Early Stopping
    - 4.6 Role of Independent Data Monitoring Committee (IDMC)
5. Data Analysis Considerations
    - 5.1 Prespecification of the Analysis
    - 5.2 Analysis Sets
        - 5.2.1 Full Analysis Set
        - 5.2.2 Per Protocol Set
        - 5.2.3 Roles of the Different Analysis Sets
    - 5.3 Missing Values and Outliers
    - 5.4 Data Transformation
    - 5.5 Estimation, Confidence Intervals and Hypothesis Testing
    - 5.6 Adjustment of Significance and Confidence Levels
    - 5.7 Subgroups, Interactions, and Covariates
    - 5.8 Integrity of Data and Computer Software Validity
6. Evaluation of Safety and Tolerability
    - 6.1 Scope of Evaluation
    - 6.2 Choice of Variables and Data Collection
    - 6.3 Set of Subjects to be Evaluated and Presentation of Data
    - 6.4 Statistical Evaluation
    - 6.5 Integrated Summary
7. Reporting
    - 7.1 Evaluation and Reporting
    - 7.2 Summarizing the Clinical Database
        - 7.2.1 Efficacy Data
        - 7.2.2 Safety Data

Glossary

caDSR sentinel tool (http://cadsrsentinel.nci.nih.gov/cadsrsentinel/do/logon).

To ensure the safety of patients participating in clinical trials, every clinical trial must be approved by a local institutional review board (IRB) at its inception and on an annual basis. An IRB typically consists of internal and external clinical trial experts, as well as community representatives. Every trial must have a sound scientific and ethical justification. In addition to the IRB's oversight, most randomized clinical trials must also be monitored by a DMC or a DSMB. The rules for the operation of the DSMB in the United States have been well established by the US National Institutes of Health (http://grants.nih.gov/grants/guide/notice-files/not98-084.html). The main function of such groups is to monitor patient accrual and safety and treatment efficacy for each clinical trial. For double-blinded studies, the DMC can ask the study statistician to provide unblinded data in a closed session. The integrity of each study can be preserved through independent reviews by the DMC. Recommendations for early stopping due to toxicity, futility, or efficacy can be made by the DMC and communicated to the study investigators and regulatory bodies (Ellenberg *et al.*, 2003; Clemens *et al.*, 2005).

Standard statistical reports, including descriptive and summary statistics, should be provided for the review of each study on at least an annual basis. Event charts, such as a calendar event chart and an interval event chart, are useful graphical tools to track and plot multiple timed event

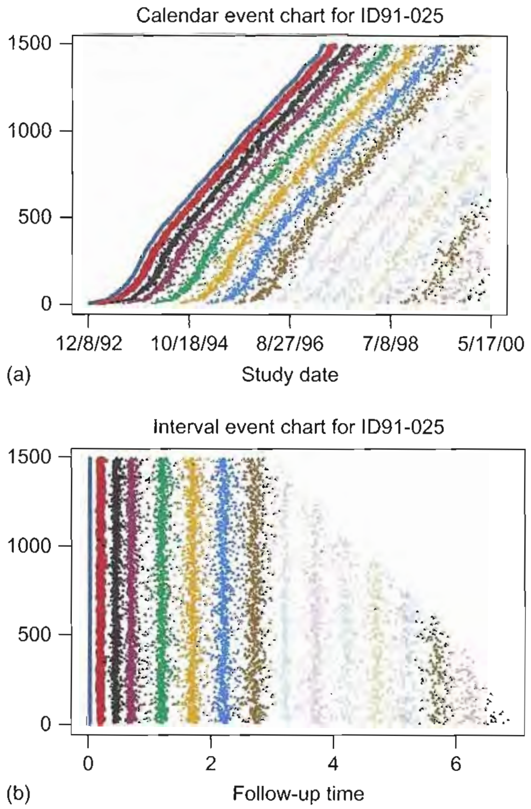

Figure 2 Event charts for monitoring clinical trial conduct of the Lung Intergroup Trial: (a) calendar event chart and (b) interval event chart.

data at the individual level (Lee *et al.*, 2000). They are complementary to the commonly used Kaplan–Meier survival plots, which provide a summary for the grouped data. They are highly effective for monitoring patient accrual and for scheduling in the conduct of clinical trials. Examples of calendar and interval event charts from the Lung Intergroup Trial (Lippman *et al.*, 2001) are shown in Figure 2. Event charts can also be very useful for assessing covariate effects.

The results from a clinical trial should be published upon completion of the trial, regardless of whether the results are positive or negative. To standardize the report of randomized clinical trials in the literature, the consolidated standards of reporting trials (CONSORT) group proposed that all clinical trials should be summarized according to CONSORT guidelines (Altman, 1996; Moher *et al.*, 2001). The number of patients registered and randomized into a trial, as well as the follow-up status over time can be easily shown in a CONSORT diagram. For example, the conduct of the Lung Intergroup Trial is summarized in such a diagram in Figure 3.

Finally, there are many useful resources available on the Internet. Some useful links are given in Table 6.

Table 6 Clinical trial resources on the web.

Name	Link	Comments
Examples of free web-based tools		
Accelerated Titration Designs for Phase I Clinical Trials	linus.nci.nih.gov/~brb/Methodologic.htm	Excel template and S-PLUS programme
BRB Array Tools for visualization and analysis of microarray data	linus.nci.nih.gov/BRB-ArrayTools.html	Windows-Excel programme
Bryant and Day Design	www.biostats.upci.pitt.edu/biostats/ClinicalStudyDesign/Phase2BryantDay.html	Phase II design considers both response and toxicity as end points
CRAB/SWOG Statistical Tools	www.crab.org/Statistools.asp	Web-based tools for the design and analysis of data from common distributions
Dartmouth Biostatistics	biostat.hitchcock.org/BSR/default.asp	Resources for clinical trials and software tools
Johns Hopkins Biostatistics	www.biostat.jhsph.edu/research/software.shtml	Randomization programmes, CRM, optimal two-stage design, sample size, and power calculation. DOS/Windows programmes
M D Anderson Biostatistics Software Download	biostatistics.mdanderson.org/SoftwareDownload	General statistics/biostatistics tools and programme for various trial designs. DOS/Windows/S-Plus/R programmes
MGH Statistical Software	hedwig.mgh.harvard.edu/biostatistics/software.php	Sample size calculation, computing sequential boundaries, etc. DOS/Windows Programmes
MRC Biostatistics	www.mrc-bsu.cam.ac.uk	Resources for clinical trials and software tools
Simon's two-stage phase II designs	linus.nci.nih.gov/~brb/Opt.htm	DOS programme
StatLib	lib.stat.cmu.edu	Numerous general statistics/biostatistics tools. DOS/Windows/S-Plus/R/MatLab/Stata programmes
STPLAN	biostatistics.mdanderson.org/SoftwareDownload	Sample size and power calculation programme
UCLA Power Calculator	calculators.stat.ucla.edu/powercalc	Sample size calculation for common problems
UPCI	www.biostats.upci.pitt.edu/biostats/ClinicalStudyDesign/	Clinical trial design resources, applets for phase I and phase II designs
Examples of Commercially Available Software Tools		
ADDPLAN	www.addplan.org	Adaptive design
ADEPT	www.rdg.ac.uk/mps/mps_home/software/adept/overview.htm	Bayesian design and conduct of phase I dose escalation studies based on Bayesian decision procedures SAS/AF application for Windows
East	www.cytel.com	Sequential design
N-Query	www.statsol.ie/html/nquery/nquery_home.html	Sample size and power calculation
PASS	www.ncss.com/passsequence.html	Sample size and power calculation
PEST 4	www.rdg.ac.uk/mps/mps_home/software/pest4/pest4.htm	An easy-to-use, competitively priced software package for the design, simulation, interim monitoring, and analysis of group sequential clinical trials
S + SeqTrial	www.statsci.com/products/seqtrial/default.asp	Sequential design. S-PLUS

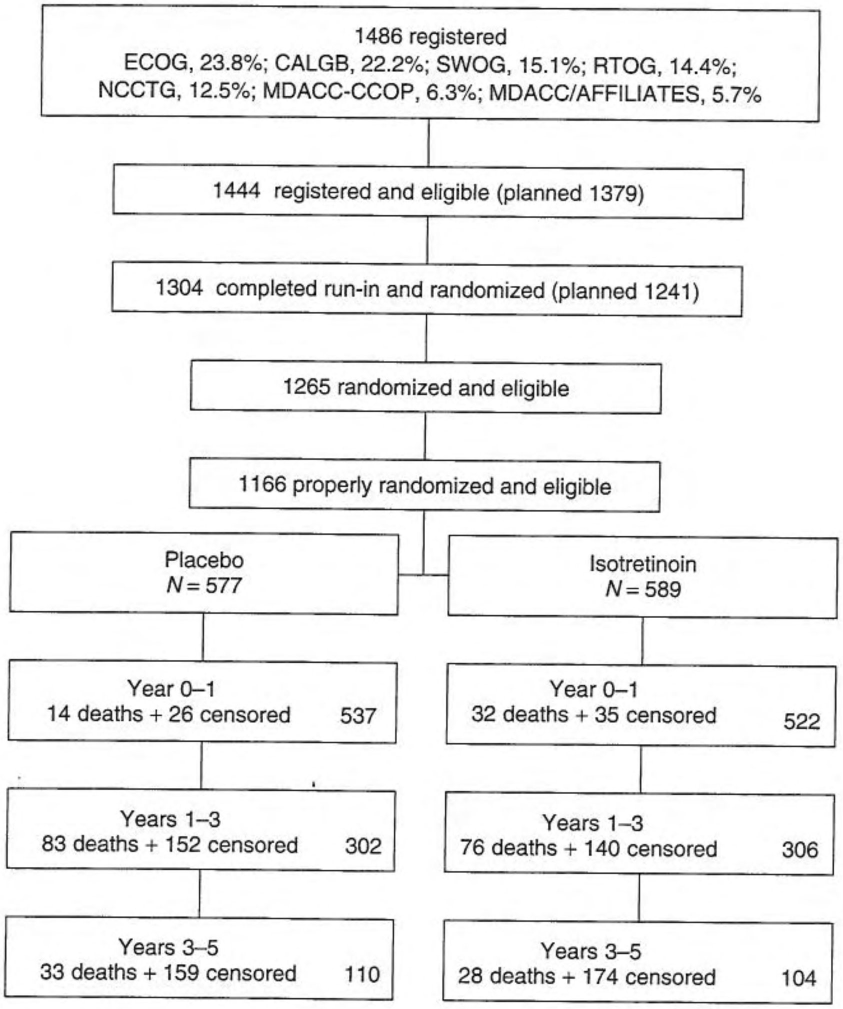
1486 registered
ECOG, 23.8%; CALGB, 22.2%; SWOG, 15.1%; RTOG, 14.4%;
NCCTG, 12.5%; MDACC-CCOP, 6.3%; MDACC/AFFILIATES, 5.7%
1444 registered and eligible (planned 1379)
1304 completed run-in and randomized (planned 1241)
1265 randomized and eligible
1166 properly randomized and eligible
Placebo
N = 577
Isotretinoin
N = 589
Year 0–1
14 deaths + 26 censored
537
Year 0–1
32 deaths + 35 censored
522
Years 1–3
83 deaths + 152 censored
302
Years 1–3
76 deaths + 140 censored
306
Years 3–5
33 deaths + 159 censored
110
Years 3–5
28 deaths + 174 censored
104

Figure 3 CONSORT diagram for reporting the results of the Lung Intergroup Trial.

ACKNOWLEDGEMENTS

The work was supported in part by the grants from the US National Cancer Institute CA16672 and CA97007. The editorial assistance from Ms Lee Ann Chastain is much appreciated.

REFERENCES

Ahn, C. (1998). An evaluation of phase I cancer clinical trial designs. *Statistics in Medicine*, **17**, 1537–1549.

Allison, D. B., *et al.* (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews. Genetics*, **7**, 55–65.

Altman, D. G. (1996). Better reporting of randomised controlled trials: the CONSORT statement. *British Medical Journal*, **313**, 570–571.

Arteaga, C. L. and Baselga, J. (2003). Clinical trial design and end points for epidermal growth factor receptor-targeted therapies: implications for drug development and practice. *Clinical Cancer Research*, **9**, 1579–1589.

Babb, J., *et al.* (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, **17**, 1103–1120.

Baquet, C. R., *et al.* (2006). Recruitment and participation in clinical trials: socio-demographic, rural/urban, and health care access predictors. *Cancer Detection and Prevention*, **30**, 24–33.

Berry, D. A. (2005). Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clinical Trials*, **2**, 295–300. discussion 301–304, 364–378.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews. Drug Discovery*, **5**, 27–36.

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, **51**, 1372–1383.

Capra, W. B. (2004). Comparing the power of the discontinuation design to that of the classic randomized design on time-to-event endpoints. *Controlled Clinical Trials*, **25**, 168–177.

Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*, **16**, 2701–2711.

Chou, T. Y., *et al.* (2005). Mutation in the tyrosine kinase domain of epidermal growth factor receptor is a predictive and prognostic factor for gefitinib treatment in patients with non–small-cell lung cancer. *Clinical Cancer Research*, **11**, 3750–3757.

Christen, W. G., *et al.* (2000). Design of Physicians' Health Study II – a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Annals of Epidemiology*, **10**, 125–134.

Clemens, F., *et al.* (2005). Data monitoring in randomized controlled trials: surveys of recent practice and policies. *Clinical Trials*, **2**, 22–33.

Collins, J. M., *et al.* (1990). Pharmacologically guided phase I clinical trials based upon preclinical drug development. *Journal of the National Cancer Institute*, **82**, 1321–1326.

Comis, R. L., *et al.* (2003). Public attitudes toward participation in cancer clinical trials. *Journal of Clinical Oncology*, **21**, 830–835.

Crowley, J. and Ankerst, D. P. (2006). *Handbook of Statistics in Clinical Oncology*, 2nd edition. Chapman & Hall/CRC, Boca Raton.

Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Ellenberg, S. S., *et al.* (2003). *Data Monitoring Committees in Clinical Trials: A Practical Perspective*, John Wiley & Sons, Chichester.

Ensign, L. G., *et al.* (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*, **13**, 1727–1736.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics*, **38**, 143–151.

Fox, E., *et al.* (2002). Clinical trial design for target-based therapy. *The Oncologist*, **7**, 401–409.

Frei, E. III *et al.* (1958). A comparative study of two regimens of combination chemotherapy in acute leukemia. *Blood*, **13**, 1126–1148.

Freidlin, B. and Simon, R. (2005a). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, **11**, 7872–7878.

Freidlin, B. and Simon, R. (2005b). Evaluation of randomized discontinuation design. *Journal of Clinical Oncology*, **23**, 5094–5098.

Friedman, L. M., *et al.* (1998). *Fundamentals of Clinical Trials*, Springer-Verlag, New York.

Fuse, E., *et al.* (1994). Application of pharmacokinetically guided dose escalation with respect to cell cycle phase specificity. *Journal of the National Cancer Institute*, **86**, 989–996.

Garrett-Mayer, E. (2006). The continual reassessment method for dose-finding studies: a tutorial. *Clinical Trials*, **3**, 57–71.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*, **13**, 346–353.

Gehan, E. A. (1997). The scientific basis of clinical trials: statistical aspects. *Clinical Cancer Research*, **3**, 2587–2590.

Geller, N. L. (1987). Planned interim analysis and its role in cancer clinical trials. *Journal of Clinical Oncology*, **5**, 1485–1490.

Geller, N. L. and Pocock, S. J. (1987). Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics*, **43**, 213–223.

Ginsburg, G. S. and Angrist, M. (2006). The future may be closer than you think: a response from the personalized medicine coalition to the Royal Society's report on personalized medicine. *Personalized Medicine*, **3**, 119–123.

Goodman, S. N., *et al.* (1995). Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*, **14**, 1149–1161.

Gray, R., *et al.* (2006). Phase II clinical trial design: methods in translational research from the genitourinary committee at the Eastern Cooperative Oncology Group. *Clinical Cancer Research*, **12**, 1966–1969.

Green, S., *et al.* (2003). *Clinical Trials in Oncology*, 2nd edition. Chapman & Hall, Boca Raton.

Heinonen, O. P., *et al.* (1994). The Alpha-Tocopherol, Beta-Carotene Lung Cancer Prevention Study: design, methods, participant characteristics, and compliance. *Annals of Epidemiology*, **4**, 1–10.

Hennekens, C. H. and Eberlein, K. (1985). A randomized trial of aspirin and beta-carotene among U.S. physicians. *Preventive Medicine*, **14**, 165–168.

Hu, J., *et al.* (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics*, **21**, 3264–3272.

ICH. (1999). ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Statistics in Medicine*, **18**, 1905–1942.

Inoue, L. Y. T., *et al.* (2002). Seamlessly expanding a randomized phase II trial to phase III. *Biometrics*, **58**, 823–831.

Jain, K. K. (2005). Personalised medicine for cancer: from drug development into clinical practice. *Expert Opinion on Pharmacotherapy*, **6**, 1463–1476.

Jemal, A., *et al.* (2006). Cancer statistics, 2006. *CA: A Cancer Journal for Clinicians*, **56**, 106–130.

Julious, S. A. (2004). Tutorial in biostatistics: sample sizes for clinical trials with normal data. *Statistics in Medicine*, **23**, 1921–1986.

Julious, S. A., *et al.* (1999). Estimating sample sizes for continuous, binary, and ordinal outcomes in paired comparisons: practical hints. *Journal of Biopharmaceutical Statistics*, **9**, 241–251.

Julious, S. A. and Patterson, S. D. (2004). Sample sizes for estimation in clinical research. *Pharmaceutical Statistics*, **3**, 213–215.

Jung, S. H., *et al.* (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, **6**, 157–169.

Korn, E. L., *et al.* (1994). A comparison of two phase I trial designs. *Statistics in Medicine*, **13**, 1799–1806.

Korn, E. L., *et al.* (2001). Clinical trial designs for cytostatic agents: are new approaches needed? *Journal of Clinical Oncology*, **19**, 265–272.

Lee, J. J. and Feng, L. (2005). Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology*, **23**, 4450–4457.

Lee, M. L. and Whitmore, G. A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, **21**, 3543–3570.

Lee, J. J., *et al.* (2000). Extensions and applications of event charts. *The American Statistician*, **54**, 63–70.

Lee, J. J., *et al.* (2001). Design considerations for efficient prostate cancer chemoprevention trials. *Urology*, **57**, 205–212.

Lewis, J. A. (1999). Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in Medicine*, **18**, 1903–1904.

Lewis, J., *et al.* (2001). The impact of the international guideline entitled Statistical Principles for Clinical Trials (ICH E9). *Statistics in Medicine*, **20**, 2549–2560.

Lin, Y. and Shih, W. J. (2001). Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics*, **2**, 203–215.

Lippman, S. M., *et al.* (2001). Randomized phase III intergroup trial of isotretinoin to prevent second primary tumors in stage I non-small-cell lung cancer. *Journal of the National Cancer Institute*, **93**, 605–618.

Liu, P. Y., *et al.* (1999). False positive rates of randomized phase II designs. *Controlled Clinical Trials*, **20**, 343–352.

Lynch, T. J., *et al.* (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, **350**, 2129–2139.

Maitournam, A. and Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*, **24**, 329–339.

Mariani, L. and Marubini, E. (1996). Design and analysis of phase II cancer trials: a review of statistical methods and guidelines for medical researchers. *International Statistical Review. Revue Internationale de Statistique*, **64**, 61–88.

McShane, L. M., *et al.* (2005). Reporting recommendations for tumor marker prognostic studies (REMARK). *Journal of the National Cancer Institute*, **97**, 1180–1184.

Michiels, S., *et al.* (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Moher, D., *et al.* (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *The Journal of the American Medical Association*, **285**, 1987–1991.

Nottage, M. and Siu, L. L. (2002). Principles of clinical trial design. *Journal of Clinical Oncology*, **20**, 42S–46S.

Nurgat, Z. A., *et al.* (2005). Patient motivations surrounding participation in phase I and phase II clinical trials of cancer chemotherapy. *British Journal of Cancer*, **92**, 1001–1005.

O'Quigley, J. and Zohar, S. (2006). Experimental designs for phase I and phase I/II dose-finding studies. *British Journal of Cancer*, **94**, 609–613.

O'Quigley, J., *et al.* (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, **46**, 33–48.

Paez, J. G., *et al.* (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, **304**, 1497–1500.

Page, G. P., *et al.* (2006). The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics*, **7**, 84.

Park, J. W., *et al.* (2004). Rationale for biomarkers and surrogate end points in mechanism-driven oncology drug development. *Clinical Cancer Research*, **10**, 3885–3896.

Parulekar, W. R. and Eisenhauer, E. A. (2004). Phase I trial design for solid tumor studies of targeted, non-cytotoxic agents: theory and practice. *Journal of the National Cancer Institute*, **96**, 990–997.

Penny, M. A. and Mchale, D. (2005). Pharmacogenomics and the drug discovery pipeline: when should it be implemented? *American Journal of Pharmacogenomics: Genomics-Related Research in Drug Development and Clinical Practice*, **5**, 53–62.

Phillips, A. and Haudiquet, V. (2003). ICH E9 guideline 'Statistical Principles for Clinical Trials': a case study. *Statistics in Medicine*, **22**, 1–11. discussion 13–17.

Piantadosi, S. (2005). *Clinical Trials: A Methodologic Perspective*, Wiley-Interscience, Hoboken.

Pocock, S. J. (1983). *Clinical Trials: A Practical Approach*, Wiley & Sons, Chichester.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.

Pusztai, L. and Hess, K. R. (2004). Clinical trial design for microarray predictive marker discovery and assessment. *Annals of Oncology*, **15**, 1731–1737.

Rogatko, A., *et al.* (2005). New paradigm in dose-finding trials: patient-specific dosing and beyond phase I. *Clinical Cancer Research*, **11**, 5342–5346.

Rosenberger, W. F. and Lachin, J. M. (2002). *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons, New York.

Rosner, G. L., *et al.* (2002). Randomized discontinuation design: application to cytostatic antineoplastic agents. *Journal of Clinical Oncology*, **20**, 4478–4484.

Scher, H. I. and Heller, G. (2002). Picking the winners in a sea of plenty. *Clinical Cancer Research*, **8**, 400–404.

Schiller, J. H. (2004). Clinical trial design issues in the era of targeted therapies. *Clinical Cancer Research*, **10**, 4281s–4282s.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, **10**, 1–10.

Simon, R. (2005). Roadmap for developing and validating therapeutically relevant genomic classifiers. *Journal of Clinical Oncology*, **23**, 7332–7341.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, **10**, 6759–6763.

Simon, R., *et al.* (1985). Randomized phase II clinical trials. *Cancer Treatment Reports*, **69**, 1375–1381.

Simon, R., *et al.* (1997). Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute*, **89**, 1138–1147.

Smith, T. L., *et al.* (1996). Design and results of phase I cancer clinical trials: three-year experience at M.D. Anderson Cancer Center. *Journal of Clinical Oncology*, **14**, 287–295.

Stadler, W. M., *et al.* (2005). Successful implementation of the randomized discontinuation trial design: an application to the study of the putative antiangiogenic agent carboxyaminoimidazole in renal cell carcinoma–CALGB 69901. *Journal of Clinical Oncology*, **23**, 3726–3732; Erratum in: (2005). *Journal of Clinical Oncology*, **23**, 4808.

Stern, A. G. (2003). COXIBs: Interpreting the swell of phase IV data. *Journal of Clinical Rheumatology: Practical Reports on Rheumatic and Musculoskeletal Diseases*, **9**, 337–339.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, **45**, 925–937.

Tannock, I. F. (1995). The recruitment of patients into clinical trials. *British Journal of Cancer*, **71**, 1134–1135.

Teicher, B. A. and Andrews, P. A. (2004). *Anticancer Drug Development Guide: Preclinical Screening, Clinical Trials, and Approval*, Humana Press, Totowa.

Temple, R. J. (2005). Enrichment designs: efficiency in development of cancer treatments. *Journal of Clinical Oncology*, **23**, 4838–4839.

Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, **60**, 684–693.

Thall, P. F., *et al.* (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, **22**, 763–780.

Therasse, P., *et al.* (2000). New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, **92**, 205–216.

Therasse, P., *et al.* (2006). RECIST revisited: a review of validation studies on tumour assessment. *European Journal of Cancer*, **42**, 1031–1039.

Wang, S. J. and Chen, J. J. (2004). Sample size for identifying differentially expressed genes in microarray experiments. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **11**, 714–726.

Wang, Y. G., *et al.* (2005). Bayesian designs with frequentist and Bayesian error rate considerations. *Statistical Methods in Medical Research*, **14**, 445–456.

FURTHER READING

Khuri, F. R., *et al.* (2006). Randomized phase III trial of low-dose isotretinoin for prevention of second primary tumors in stage I and II head and neck cancer patients. *Journal of the National Cancer Institute*, **98**, 441–450.

Korn, E. L. and Simon, R. (1993). Using the tolerable-dose diagram in the design of phase I combination chemotherapy trials. *Journal of Clinical Oncology*, **11**, 794–801.

Piantadosi, S., *et al.* (1998). Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemotherapy and Pharmacology*, **41**, 429–436.

Stallard, N. (2003). Decision-theoretic designs for phase II clinical trials allowing for competing studies. *Biometrics*, **59**, 402–409.