Speeding up the Evaluation of New Agents in Cancer

Mahesh K. B. Parmar, Friederike M.-S. Barthel, Matthew Sydes, Ruth Langley, Rick Kaplan, Elizabeth Eisenhauer, Mark Brady, Nicholas James, Michael A. Bookman, Ann-Marie Swart, Wendi Qian, Patrick Royston

Despite both the increase in basic biologic knowledge and the fact that many new agents have reached various stages of development during the last 10 years, the number of new treatments that have been approved for patients has not increased as expected. We propose the multi-arm, multi-stage trial design as a way to evaluate treatments faster and more efficiently than current standard trial designs. By using intermediate outcomes and testing a number of new agents (and combinations) simultaneously, the new design requires fewer patients. Three trials using this methodology are presented.

J Natl Cancer Inst 2008;100:1204–1214

During the last 10 years, there has been a huge increase in the understanding of many diseases, based on a revolution in the molecular sciences (1,2). This knowledge has inevitably fueled considerable hope in the potential to cure many serious diseases, such as cancer, HIV/AIDS, and heart disease. In cancer, for example, there has been a dramatic and unprecedented increase in the number of potential new anticancer therapies in recent years. In 2005, it was estimated that 1994 anticancer agents were in development, including 195, 389, and 122 in clinical phases 1, 2, and 3, respectively (3). Many of these agents result from advances in our understanding of cell biology, in particular, intracellular signaling pathways, growth factors and their receptors, and increased knowledge of the human genome. A substantial proportion of the agents are aimed at the same few molecular targets, such as the epidermal growth factor receptor and vascular endothelial growth factor receptor (4).

However, in a report in March 2004, the US Food and Drug Administration (FDA) (5) identified a slowdown, rather than the expected acceleration, in innovative medical therapies being approved and reaching patients. Three factors have been highlighted as being involved in this downturn: 1) the high costs of bringing a new product to market, which is estimated to be of the order of US $1.2 billion or more (3); 2) the fact that most new treatments are not effective—the FDA has estimated that only approximately 8% of therapies entering phase 1 trials reach the market (5); and 3) changes in the regulatory requirements for licensing approval.

A consequence of this slowdown in approvals is the concern that the hoped-for advances in improving survival and quality of life in many major diseases may not materialize. This downturn has happened even though biomedical research spending has more than doubled in real terms in the private sector globally over the last 10 years (5). There have also been increases in public sector research funding internationally. For example, in the United Kingdom, spending from all sources, private and public, on biomedical research and development increased by 14% in real terms between 1994 and 2000 (6).

To respond to this slowdown, the FDA has called for new additions to the "product-development toolkit" to achieve reliable results more rapidly (5). In this commentary, we present one approach that addresses this need.

Principles

There are many steps in the process of developing and evaluating new therapies. Here, we discuss some critical components of this process and provide an impetus for an alternative approach.

Acknowledge that Phase 2 Trials, as Currently Conducted, Are Not a Sufficiently Good Screen for Identifying Potentially Effective Therapies

The very large proportion of the recent cost increases in drug development and testing [estimated as a 55% increase in the last 5 years (5)] are accumulated during the phase 2 and 3 components of the process. Phase 3 trials represent 65%–75% of the costs of the clinical phase portion (7). A critical decision point is the selection of therapies to enter larger-scale randomized testing in phase 3 trials. Phase 2 trials are usually designed as a "screen" to assess whether there is sufficient therapeutic activity and an acceptable toxicity profile to warrant further testing and development in larger scale randomized phase 3 trials. There is, however, a distinction between phase 2 trials that use the new drug as a single agent and those that use the new drug in combination with current routine therapies. Although in both types of phase 2 trials the primary concerns are safety and toxicity, the two types of trials differ in their aims of assessing activity. Single-agent phase 2 trials are useful in assessing whether the agent has a minimum

Affiliations of authors: MRC Clinical Trials Unit, London, UK (MKBP, MS, RL, RK, AMS, WQ, PR); Institute of Psychiatry, King's College London, London, UK (FMSB); NCIC Clinical Trials Group, Queen's University, Kingston, Ontario, Canada (EE); School of Public Health and Health Professions, University at Buffalo, Buffalo, NY (MB); CR-UK Institute for Cancer Studies, University of Birmingham, Birmingham, UK (NJ); Fox Chase Cancer Center, Philadelphia, PA (MAB).

Correspondence to: Mahesh K. B. Parmar, DPhil, MRC Clinical Trials Unit, 222 Euston Road, London, NW1 2DA, UK (e-mail: mp@ctu.mrc.ac.uk).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djn267

level of activity that would warrant further investigation. In contrast, in phase 2 trials of combination therapy, activity data are difficult to interpret because there will be an unquantifiable response to the underlying therapy and no randomized comparison is made. Furthermore, the relationship between any potential improvements in response rate and longer-term outcome measures, such as overall survival, remains unclear. One of the major difficulties in assessing the need for a randomized phase 3 trial is the relatively poor evidence that is provided by the noncomparative nature of phase 2 combination therapy trials. Although randomized controlled phase 2 trials have been proposed (8), these designs do not generally provide robust or reliable evidence on which to base a decision regarding further testing because there is no direct comparison between the new therapy and the control group.

Accept that the Size of the Effect of Most New Therapies on Important Outcome Measures, Such As Overall and Disease-Specific Survival, Is Usually Modest

During the last 20 years, it has become apparent that improvements in survival provided by new cancer agents, when added to standard care, are generally modest (9,10). Two examples of this are as follows. In the first-line treatment of patients with metastatic colorectal cancer, the addition of the drug bevacizumab to oxaliplatin-based chemotherapy improved median survival from 19.9 to 21.3 months (hazard ratio [HR] = 0.89, 95% confidence interval [CI] = 0.7 to 1.03) (11). In patients with mesothelioma, the drug pemtrexed improved median survival from 9.3 to 12.1 months (HR = 0.77, estimated 95% CI = 0.61 to 0.96) when added to cisplatin-based therapy (12).

Acknowledge That Only a Small Proportion of New Therapies Will Prove To Be Better Than Current Standard Therapies

The FDA report (5) emphasizes that only 8% of new drugs that enter phase 1 trials actually reach the market. In cancer, Roberts et al. (9) found that of the 208 antineoplastic agents brought into clinical trials from 1975 to 1994, only 29 (14%) ultimately received FDA marketing approval. In a different setting, a review of the randomized controlled trials conducted by the Children's Oncology Group in the United States (13) showed that in terms of the trend of an effect, new treatments are as likely to appear inferior to standard treatments as they are to appear superior.

Kola and Landis (14) reviewed the success rates of new agents for a range of diseases from the top 10 pharmaceutical companies during the period 1991–2000. Their results suggest that the success rate in phase 2, which is estimated as the proportion of new agents going on to be tested in phase 3, is reasonably independent of disease and is estimated to be between 30% and 40%. The success of phase 3 trials, which is defined as a positive result from the trial, varied more across diseases, from 40% in oncology to 75% in cardiovascular disease. On average, across all diseases about 50% of phase 3 trials are successful and lead to a licensing application. A 40%–50% success rate for randomized phase 3 trials in cancer may be considered too low because it is no better than tossing a coin. The final hurdle for new agents is the licensing stage, and on average, 70% of agents that made a licensing application received one.

A Partial Solution

Consideration of the first two principles given above has led to a proposal in some instances to bypass traditional phase 2 trials and to conduct large phase 3 trials as early as possible during the development and testing of new agents, often including many thousands of patients to reliably detect modest differences. This approach is increasingly being used by both the pharmaceutical industry and the academic sector. For example, two large-scale randomized phase 3 studies testing the addition of gefitinib (Iressa) to chemotherapy in advanced non–small cell lung cancer (15–17) were conducted with 1093 and 1037, patients recruited. These two trials were initiated after two phase 2 trials of single-agent gefitinib showed the activity of this agent in more advanced stages of the disease (17,18). None of the three randomized phase 3 trials showed an improvement in overall survival with gefitinib, despite having large numbers of events for the primary outcome.

Many thousands of large-scale trials will be required to test all the potential new agents and combinations of them with other drugs. This approach is clearly unrealistic in a reasonable time frame. Therefore, although this solution may provide reliable results about the value of a therapy in a particular setting, it does not provide an appropriate strategy to respond to the problem. In fact, performing large numbers of large-scale trials could actually exacerbate the problem because such trials can take up a large proportion of the pool of patients with the disease and prevent potentially better agents from being tested.

A New Strategy: Multi-Arm Multi-Stage Design

We therefore need other strategies in our toolkit to speed up the process of getting reliable answers. A strategy may be considered useful if it can satisfy the following principles: 1) it is better than separate single-arm phase 2 trials in deciding whether to continue testing a new treatment; 2) it will test many new promising treatments at the same time so that the probability of finding a successful new treatment is increased; 3) it has the potential to discontinue unpromising arms quickly and reliably; and 4) it bases major decisions on randomized evidence.

One approach that addresses all of the above principles is the multi-arm multi-stage (MAMS) randomized trial. In this approach, several agents are assessed simultaneously against a single control group in a randomized fashion. In the early stages of the trial, each of the experimental arms is compared in a pairwise manner with the control arm using an "intermediate" outcome measure that is required to be related to the primary outcome measure but does not have to be a true "surrogate" outcome measure [for definitions of surrogacy see (19)]. Recruitment to experimental arms that do not show sufficient promise with the intermediate outcome measures is discontinued. Recruitment to the control arm and to the promising experimental arms continues until sufficient numbers of patients have been entered to assess the impact of the experimental treatments on the primary outcome measure.

A hypothetical example is a randomized trial with four experimental arms and one control arm, run in two stages (Figure 1). The intermediate and primary outcome measures are progression-free survival and overall survival, respectively. When a prespecified

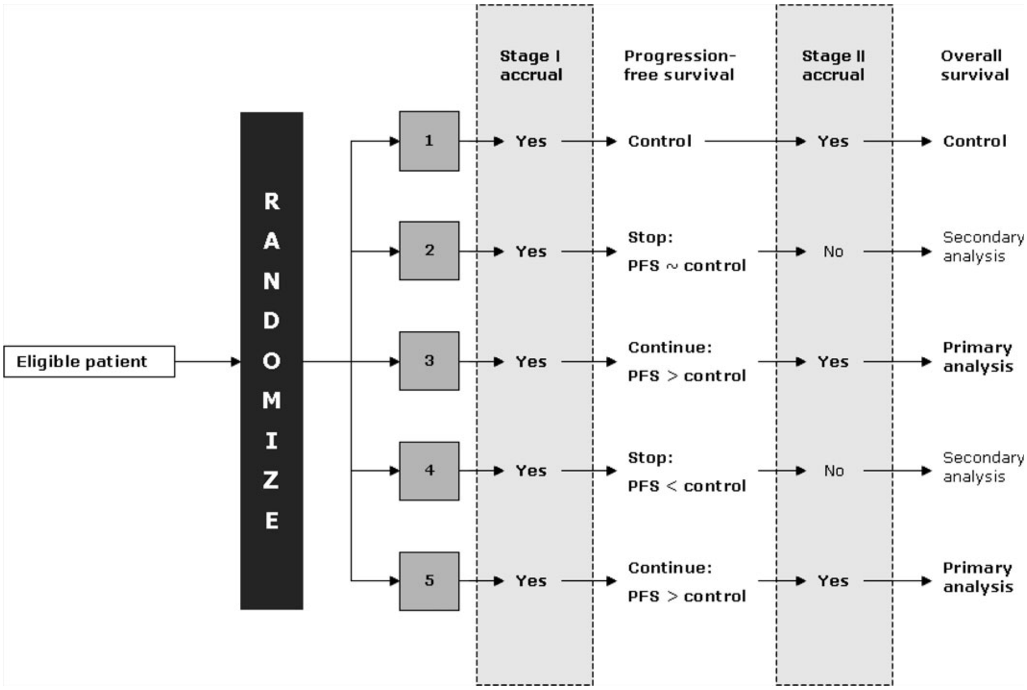

Figure 1. Hypothetical randomized trial showing a multi-arm, two-stage design. Arm 1 is the control arm and arms 2–5 are the experimental arms. At the end of stage I, each experimental arm is compared against the control arm in a pairwise manner using the intermediate outcome measure (in this case, progression-free survival). At the end of stage II, each experimental arm that has passed stage I is compared with the control arm on the primary outcome measure for the trial (primary comparison; in this case overall survival). However, secondary comparisons of experimental versus control for each arm that did not pass stage I are also performed (these comparisons will, of course, have fewer patients and events).

number of intermediate outcome events have been observed in the control arm, a pairwise comparison is made between each experimental arm and the control arm. If the observed effect size does not cross a predefined critical value, then consideration is given to not randomly assigning additional patients to that experimental arm. Accrual to the trial, however, continues while the analysis is conducted. After the analysis, patients continue to be randomly assigned to those experimental treatments that do cross the critical value and also to the control arm until the prespecified number of events on the primary outcome measure have been observed. The predefined critical value depends on four components: 1) the null hypothesis for the intermediate outcome measure (usually taken to be no difference), 2) the alternative hypothesis for the intermediate outcome measure, 3) the probability of continuing to the next stage should the null hypothesis be true, and 4) the probability of continuing to the next stage should the alternative hypothesis be true. The critical value is calculated for each stage by considering whether we can reject the null hypothesis (at the level of the probability of continuing to the next stage should the null hypothesis be true). Technical details are given in (20), and the practical specification of these parameters is displayed in the examples below.

A general explanation of an intermediate outcome measure used in this way is as follows. If there is no effect on the intermediate outcome measure (ie, if the null hypothesis is true), then it is very likely that there will be no effect on the primary outcome measure. The intermediate outcome measure is therefore required to have high negative predictive value. However, if the alternative hypothesis is true for the intermediate outcome, this will not necessarily mean that the alternative hypothesis will be true for the primary outcome measure. There is no requirement for the intermediate outcome measure to have a high positive predictive value. In trials of cancer treatment, typical intermediate outcome measures might be progression-free survival or response to treatment and a typical primary outcome measure might be overall survival. Extension of this model to more than two stages is shown in the examples below. In the MAMS design, a randomized comparison is initiated as soon as possible, although there still remains a role for single-agent phase 2 trials to prioritize new therapies for feeding into MAMS trials (Figure 2). One of the first advantages of the MAMS design is that many new treatments are considered at once, involving fewer patients over a shorter time with reduced costs than assessing each of the agents in large-scale separate two-arm trials. The multi-arm nature also improves the likelihood of a "positive" trial. For example, if a two-arm phase 3 trial in oncology has a 40% chance of showing a "positive" result (14), and if we assume that the probability of success of each of the new experimental arms in a MAMS trial is approximately independent, then for a five-arm cancer trial with four new experimental therapies, the probability of at least one successful arm in the trial increases to 87%.

We are aware of three trials that have used the MAMS design. These are the Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial (21); a collaborative trial, GOG-182/ICON5 (23), involving the Gynecologic Oncology Group (GOG) and the International Collaborative Ovarian Neoplasm Studies Group (ICON) (22); and ICON6 (23) (Table 1).

STAMPEDE

STAMPEDE (21) is a six-arm, five-stage trial of different therapies for men who are starting hormone therapy for advanced prostate cancer. Such men will typically have disease that has spread beyond the prostate, and thus it is standard care to treat their disease systemically with hormonal therapy. Approximately 85% of patients initially respond well to such hormone therapy, but the disease progresses in virtually all patients, with a median time to progression of approximately

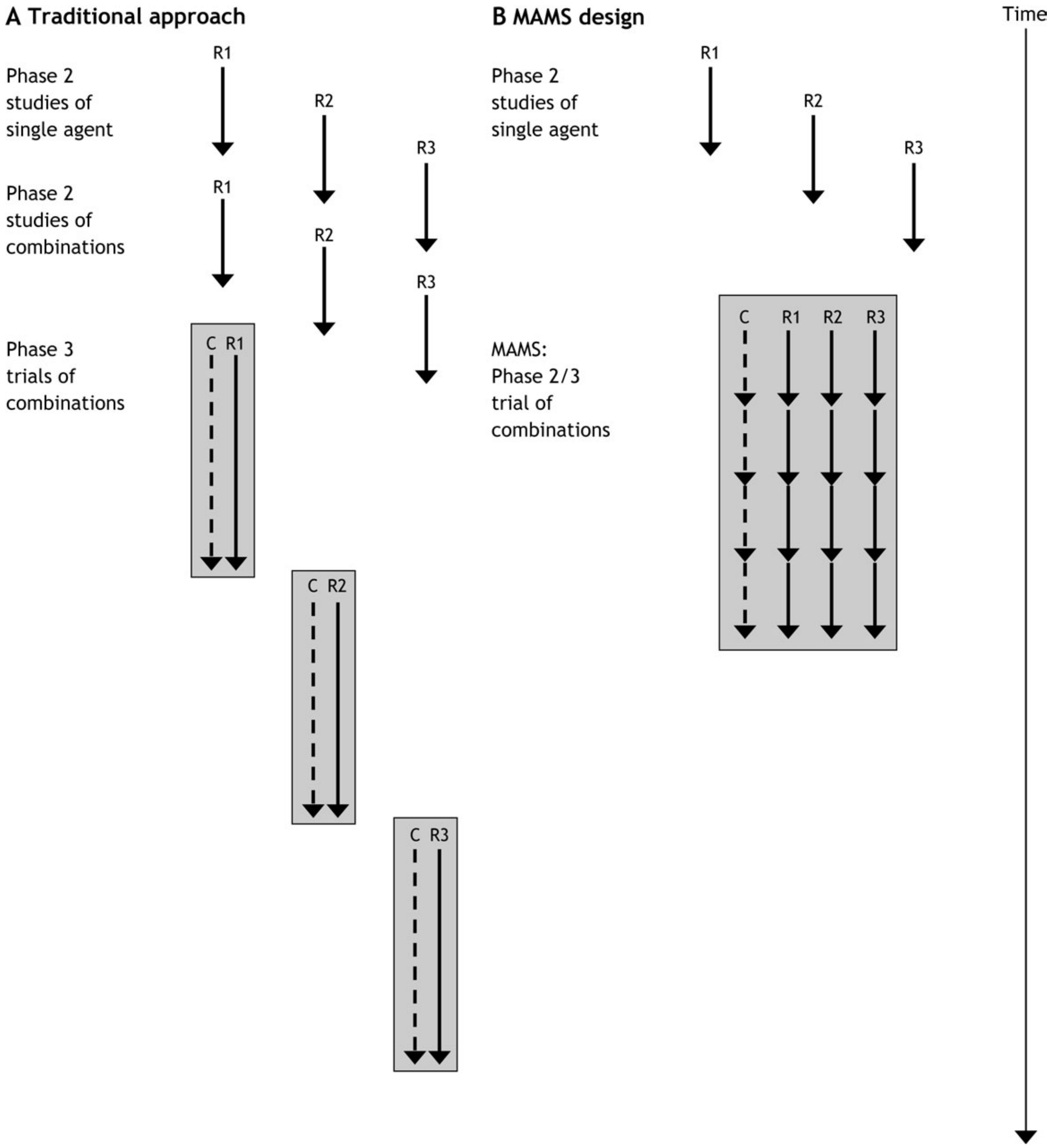

Figure 2. Where do multi-arm multi-stage (MAMS) trials fit into the phase 1, 2, and 3 setup? **A**) The traditional approach. Three new agents, R1, R2, and R3, enter and pass three single-agent single-arm phase 2 trials and also three separate single-arm combination phase 2 trials. The three combination therapies are finally compared with the control therapy in three separate randomized phase 3 trials. In this model, a total of 2100 patients are required. **B**) In the MAMS design, the single-agent single-arm phase 2 trials are followed by a single MAMS trial of all combination therapies. The MAMS model required 1300 patients in total, a saving of 800 patients. C = control arm; R1 = experimental arm R1; R2 = experimental arm R2; R3 = experimental arm R3. In these models, we assume that single-agent studies would be carried out before combination therapy studies and that phase 2 studies require only a small number of centers. Consequently, phase 2 studies of different agents may be carried out concurrently. We also assume that phase 3 trials require larger numbers of patients and a network of centers that can run only one trial in a particular group of patients at a time, and, therefore, phase 3 trials of different agents must be carried out sequentially. The MAMS design rolls the phase 2 assessment of the activity of combination therapy into the same trial as the phase 3 assessment of effectiveness.

24 months. A number of treatments, when added to hormone therapy, could potentially improve these outcomes. STAMPEDE is a trial of three of these therapies, together with some combinations of them.

In STAMPEDE, patients are randomly assigned to either the control arm or one of five experimental arms (Figure 3, A). The five stages of the trial include a pilot stage, three intermediate activity stages, and a final efficacy stage (Figure 4). The randomization ratio to the control and the five experimental arms is 2:1:1:1:1:1. The control arm is used in all the pairwise comparisons, and this imbalance in randomization facilitates a more reliable estimate of the event rates in the control arm at any given time. Moreover, for a given total number of patients to be randomly assigned to the trial, the imbalance increases the power slightly for each pairwise comparison with the control arm.

The pilot phase was planned to include 210 patients, with the aim of confirming the safety of the five experimental treatments, particularly in the two arms with treatment combinations of

Table 1. Examples of multi-arm, multi-stage trials (protocols for these trials)*

Trial name	Cancer type	Number of arms	Number of stages	Status	Number of companies involved
STAMPEDE	Hormone-naive prostate cancer	6	5	Open to accrual	3
GOG-182/ICON5	Advanced ovarian cancer	5	2	Closed to accrual—results publicly presented	3
ICON6	Relapsed ovarian cancer	3	3	Open to accrual	1

* Protocols for these trials are available from the authors on request. STAMPEDE = Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy; GOG = Gynecologic Oncology Group; ICON = International Collaborative Ovarian Neoplasm studies.

zoledronic acid plus docetaxel and zoledronic acid plus celecoxib that had not been tested before in men with prostate cancer. There was no a priori reason to suspect that any of the experimental treatments would produce unacceptable toxic effects. The three intermediate activity stages were designed to compare each experimental arm pairwise with the control arm on the intermediate outcome measure of failure-free survival (FFS, including prostate-specific antigen–defined progression). At each of these stages, the guideline

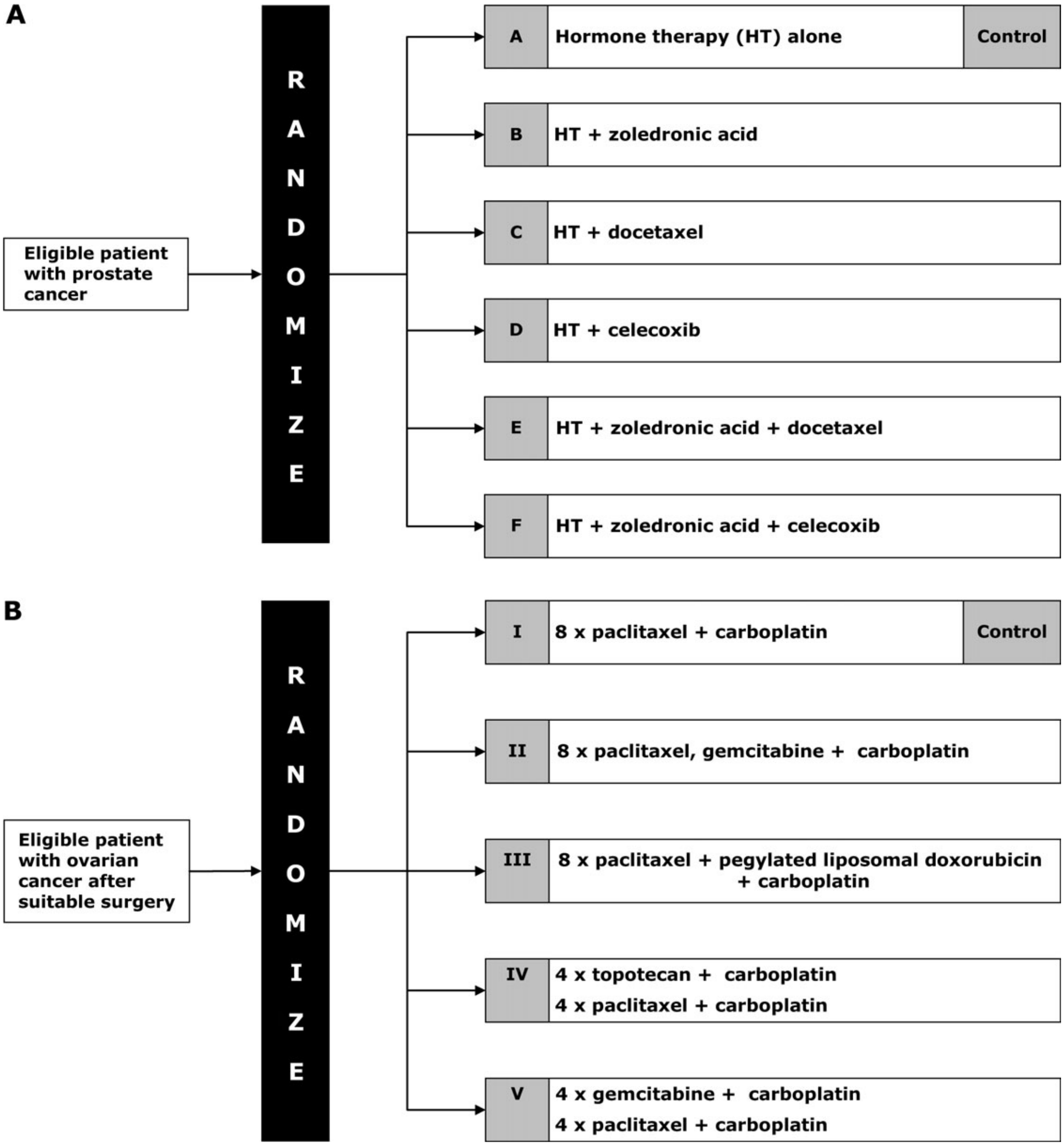
A
Eligible patient with prostate cancer
RANDOMIZE
A Hormone therapy (HT) alone Control
B HT + zoledronic acid
C HT + docetaxel
D HT + celecoxib
E HT + zoledronic acid + docetaxel
F HT + zoledronic acid + celecoxib
B
Eligible patient with ovarian cancer after suitable surgery
RANDOMIZE
I 8 x paclitaxel + carboplatin Control
II 8 x paclitaxel, gemcitabine + carboplatin
III 8 x paclitaxel + pegylated liposomal doxorubicin + carboplatin
IV 4 x topotecan + carboplatin
4 x paclitaxel + carboplatin
V 4 x gemcitabine + carboplatin
4 x paclitaxel + carboplatin

Figure 3. Two multi-arm multi-stage trials. **A**) Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial with six arms (A–F). **B**) Gynecologic Oncology Group/International Collaborative Ovarian Neoplasm Studies (GOG-182/ICON5) trial with five arms (I–V).

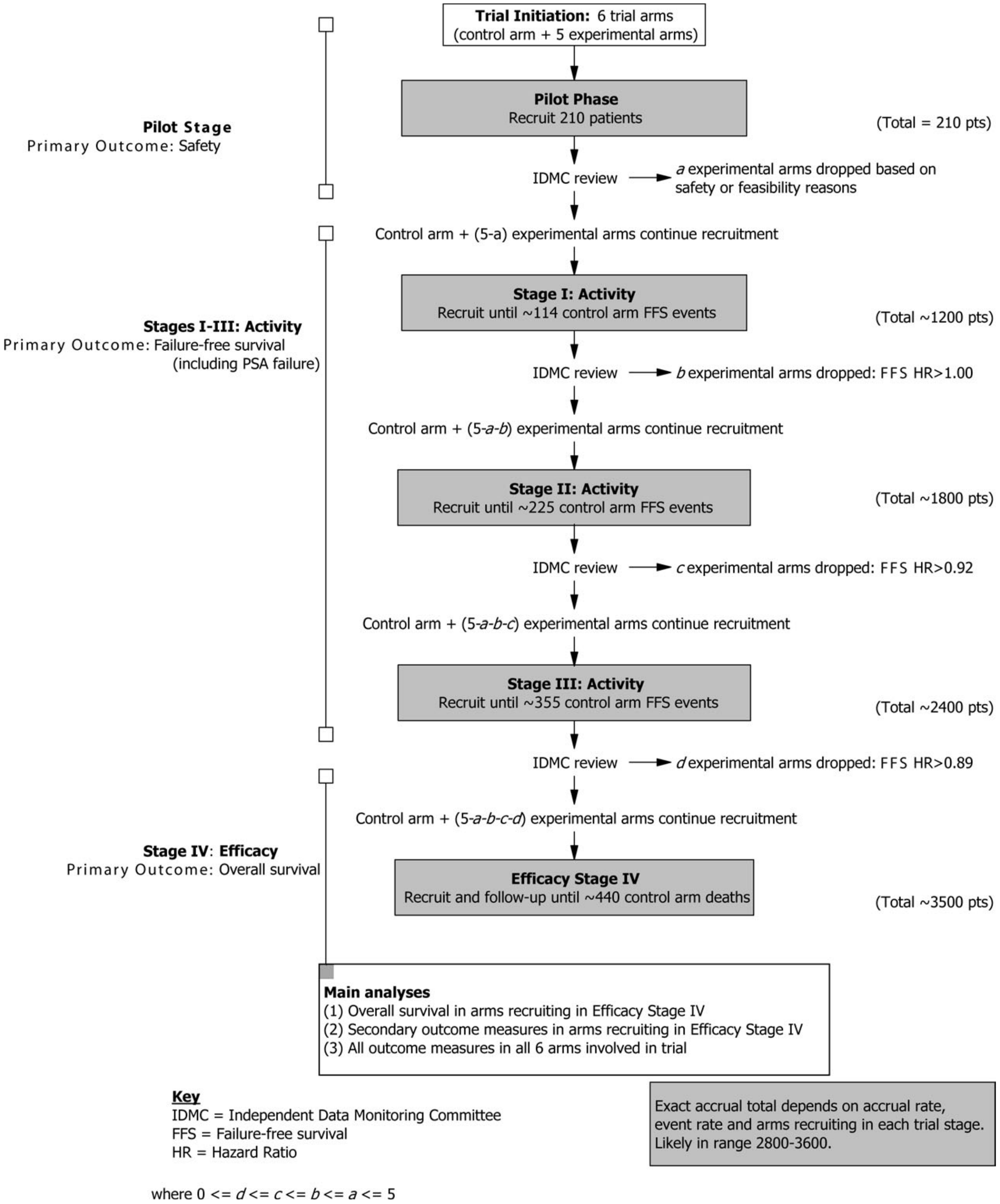
Trial Initiation: 6 trial arms
(control arm + 5 experimental arms)
Pilot Stage
Primary Outcome: Safety
Pilot Phase
Recruit 210 patients
(Total = 210 pts)
IDMC review
a experimental arms dropped based on safety or feasibility reasons
Control arm + (5-a) experimental arms continue recruitment
Stages I-III: Activity
Primary Outcome: Failure-free survival (including PSA failure)
Stage I: Activity
Recruit until ~114 control arm FFS events
(Total ~1200 pts)
IDMC review
b experimental arms dropped: FFS HR>1.00
Control arm + (5-a-b) experimental arms continue recruitment
Stage II: Activity
Recruit until ~225 control arm FFS events
(Total ~1800 pts)
IDMC review
c experimental arms dropped: FFS HR>0.92
Control arm + (5-a-b-c) experimental arms continue recruitment
Stage III: Activity
Recruit until ~355 control arm FFS events
(Total ~2400 pts)
IDMC review
d experimental arms dropped: FFS HR>0.89
Control arm + (5-a-b-c-d) experimental arms continue recruitment
Stage IV: Efficacy
Primary Outcome: Overall survival
Efficacy Stage IV
Recruit and follow-up until ~440 control arm deaths
(Total ~3500 pts)
Main analyses
(1) Overall survival in arms recruiting in Efficacy Stage IV
(2) Secondary outcome measures in arms recruiting in Efficacy Stage IV
(3) All outcome measures in all 6 arms involved in trial
Key
IDMC = Independent Data Monitoring Committee
FFS = Failure-free survival
HR = Hazard Ratio
Exact accrual total depends on accrual rate, event rate and arms recruiting in each trial stage. Likely in range 2800-3600.
where 0 <= d <= c <= b <= a <= 5

Figure 4. Five Stages of the Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial. IDMC =Independent Data Monitoring Committee; FFS =failure-free survival; HR =hazard ratio, where $0 \leq d \leq c \leq b \leq a \leq 5$.

critical value has been set for the observed HR. These critical values are 1.00, 0.92, and 0.89 for stages I, II, and III, respectively, and analyses will be performed when 115, 225, and 355 FFS events, respectively, have been observed in the control arm. The final stage has the primary outcome measure of overall survival. Key operating characteristics at each stage and overall are the error of continuing to the next stage, should the null hypothesis be true, the overall type I error, and the power (Table 2). How were the hurdles chosen? First, if an experimental arm is as effective as specified in the alternative hypothesis, then we require a high probability that it will continue to the next stage. This probability is set at 95% for stages I to III inclusive. To achieve this probability and still have an opportunity to stop an experimental arm for lack of benefit, we need to take a more “relaxed” approach to continuing to the next stage when

Table 2. Design characteristics of the STAMPEDE trial*

Stage	Primary outcome	Targeted HR (alternative hypothesis)	Critical HR†	Error	Power for targeted difference (%)	Number of events required in control arm	Expected total number of patients
Pilot	Toxicity	n/a	n/a	n/a	n/a	n/a	210
I	FFS	0.75	1.00	0.5‡	95§	115	1200
II	FFS	0.75	0.92	0.25‡	95§	225	1800
III	FFS	0.75	0.89	0.1‡	95§	355	2400
IV	OS	0.75	n/a	0.025‖	90¶	440	3200
Overall	Pairwise			0.017‖	84¶		

* HR = hazard ratio, n/a = not applicable; FFS = failure-free survival; OS = overall survival; STAMPEDE = Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy.

† The critical hazard ratio is the guideline critical value such that if the pairwise observed hazard ratio was closer to 1, then consideration would be given to discontinue further randomizations to this experimental arm.

‡ An error of this type represents the probability of continuing to the next stage when the null hypothesis (of no difference) for the intermediate outcome measure is true.

§ These values represent the probability of continuing to the next stage when the alternative hypothesis for the intermediate outcome measure is true.

‖ These errors are traditional type I errors. They represent the probability of concluding that there is a difference when the null hypothesis for the primary outcome measure is true.

¶ These values represent the "power" in the traditional sense—the probability of rejecting the null hypothesis of no difference on the primary outcome measure when the alternative hypothesis for the primary outcome measure is true.

the null hypothesis is true. An error in this direction can be considered to be "conservative." For STAMPEDE, at the end of the first stage, we have set a 50% probability of stopping each experimental arm when the null hypothesis is true. After the first stage, as the control arm events continue to accumulate and the information in the trial increases, this probability can be reduced. Thus, at the end of the second stage, the probability of continuing when the null hypothesis is true is reduced to 25%, and at the third stage it is reduced further, to 10%. The power at the end of stage IV for the outcome of overall survival is set at the traditional 90%, with a (one-sided) type I error of 2.5%. Overall, across all stages, each pairwise comparison retains good power of 84%, with an overall type I error of 1.7%. The boundaries and probabilities of stopping, assuming we were to observe an estimate from the trial exactly on the critical HR for that stage, are best displayed graphically (Figure 5).

Using a uniform distribution to model the accrual rate means that at the end of these three stages, we anticipate 1200, 1800, and 2400 patients to be randomly assigned in the entire trial. For each experimental arm, these numbers will correspond to 172, 272, and 392 patients being entered into each arm (remaining) under the assumption that five experimental arms will accrue in the first stage, four in the second, and three in the third. This trial recruited its first patient on October 17, 2005, and is anticipated to be completed within 7 years. By June 4, 2008, 582 patients had been entered. The pilot phase had been completed successfully, and all arms had been continued into the next stage.

GOG-182/ICON5

GOG-182/ICON5 is an MAMS trial with five arms and two stages. Women with advanced ovarian cancer were randomly

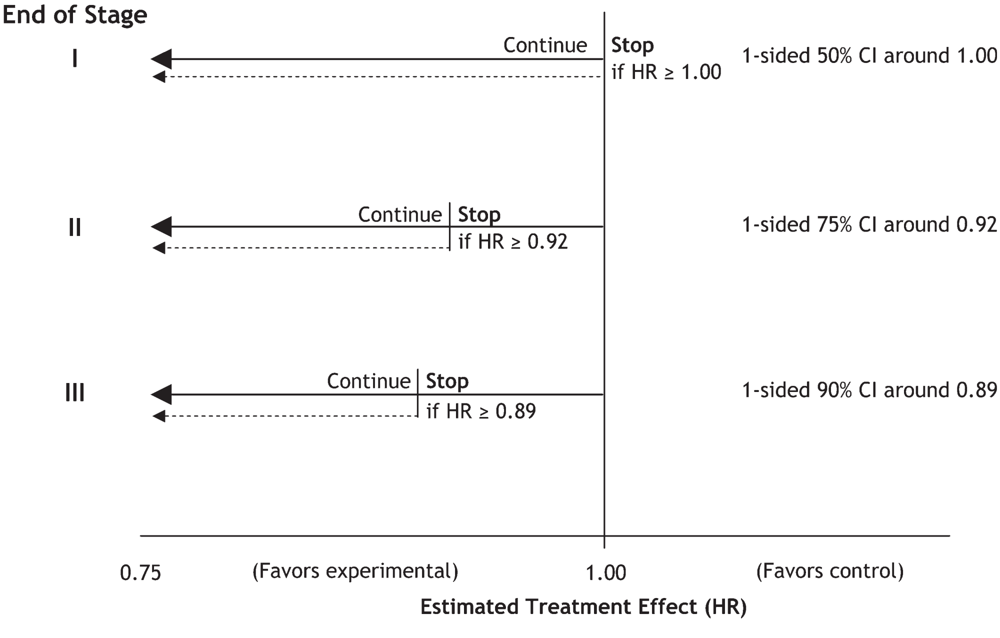

Figure 5. Stopping guidelines on the hazard ratio scale for the Systemic Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy (STAMPEDE) trial. CI = confidence interval; HR = hazard ratio; Stop = stopping of accrual (rather than termination of follow up).

Table 3. Estimated treatment hazard ratios (HRs) for progression-free survival and overall survival (ratio of experimental to control) for the first stage analysis of GOG-182/ICON5 presented to the Data Monitoring Committee in May 2004*

	Progression-free survival		Overall survival
Experimental regimen	**Crude HR (95% CI)**	**Adjusted HR†**	**Crude HR (95% CI)**
Gemcitabine triplet	0.95 (0.80 to 1.12)	0.96	0.95 (0.73 to 1.23)
Doxil triplet	0.94 (0.80 to 1.12)	0.94	1.09 (0.85 to 1.40)
Topotecan doublet	1.07 (0.90 to 1.26)	1.04	0.90 (0.69 to 1.16)
Gemcitabine doublet	1.01 (0.85 to 1.19)	0.99	1.01 (0.78 to 1.30)

* CI = confidence interval; GOG = Gynecologic Oncology Group; ICON = International Collaborative Ovarian Neoplasm Studies.

† Adjusted for stage (III vs IV), primary disease site (ovary vs extraovarian), age group (<60 vs 60–74.9 vs ≥75 years) and size of stage III residual disease (≤1 vs >1 cm).

assigned to one of five different combination chemotherapy regimens, consisting of four experimental arms and one control arm (Figure 3, B). Separate pilot trials (24–26) were conducted before GOG-182/ICON5, the main aim of which was to confirm the feasibility and safety of the new combination regimens before launching a randomized controlled trial; activity was not a major outcome measure. The first stage analysis of GOG-182/ICON5, using progression-free survival, was planned when 240 progressions or deaths in the control arm had been observed. The second stage of the trial was designed to focus on overall survival. At both stages, each of the four experimental arms was to be compared in a pairwise manner with the control arm.

The trial started accruing patients on February 7, 2001, and, with an anticipated entry rate of 500 patients per year, the 240 progressions or deaths were predicted to be observed approximately 4 years into the trial. At the outset, the guideline critical value of the hazard ratio for each pairwise comparison of progression-free survival after stage I was set at 0.87 (HR < 1 favors the experimental over the control arm). Thus, if the observed HR was greater than 0.87 (ie, closer to 1.00), then the Data Monitoring Committee (DMC) should consider recommending stopping further accrual to that particular experimental arm; if HR was less than 0.87, then accrual to the arm should be continued. Assuming that the experimental regimen was truly effective (ie, that it had a real underlying HR of 0.75), then the probability that it would be observed to be better than 0.87 was 93%, with a 5% probability that the trial would continue inappropriately.

The observed accrual rate was exceptionally high, with more than 1200 patients per year being entered into the trial worldwide over 3 years. The first stage analysis was triggered in May 2004, when 3836 patients had been randomly assigned and 272 events (progressions or deaths) had been reported in the control arm. Such a fast accrual rate gave the opportunity to relax the intermediate hurdle. Thus, the DMC considered not only the hurdle of 0.87 but also the hurdle of 0.94. This additional hurdle was introduced without knowledge of the results. This change means that if an experimental regimen was truly effective (ie, had a real underlying HR of 0.75), then the probability that it would jump this new hurdle was greater than 99.9%, with a 5% probability of continuing to the next stage, should the null hypothesis be true. This conservative and small change in the hurdle had very little impact on the overall power and type I error for the trial as a whole.

The statistical report provided to the DMC presented data on PFS, toxicity, and deaths due to treatment (Table 3). Overall survival data were also presented for context, although data for this outcome were inevitably limited. In accordance with the prespecified guidelines, the DMC saw no justification to extend accrual to any of the arms and thus indicated that the trial be closed to accrual of further patients. This conclusion was endorsed by the International Steering Committee for the trial, and hence accrual was closed on September 1, 2004. The mature results on overall survival presented in June 2006 [(22), Table 4] confirm that the decision to not accrue additional patients was a good one.

The GOG-182/ICON5 trial clearly displays the practical value of the MAMS design. Unfortunately, none of the new treatment approaches showed enough potential on the intermediate outcome measure of progression-free survival to justify continuation to the second and final stage of accrual. It was more appropriate to focus resources on assessing new approaches. However, we obtained reliable answers to these four questions in 3.5 years (from start of accrual to the planned first stage analysis), which is considerably faster than we have been able to do before. The MAMS nature of the trial saved some 20 years when compared with an alternative approach of four consecutive two-arm trials each with overall survival as the primary and only outcome measure.

ICON6

ICON6 (24) is a three-arm, three-stage double-blind placebo-controlled multicenter randomized phase 3 trial for women with relapsed ovarian cancer. The three arms of ICON6 are chemotherapy alone, chemotherapy plus cediranib given during chemotherapy, and chemotherapy plus cediranib during chemotherapy and further cediranib alone for a maximum of 18 months. The

Table 4. Updated treatment hazard ratios (HRs) for progression-free and overall survival (ratio of experimental to control) for the first stage analysis of GOG-182/ICON5 presented at the American Society of Clinical Oncology in June 2006*

	Progression-free survival	Overall survival
Experimental regimen	**Crude HR (95% CI)**	**Crude HR (95% CI)**
Gemcitabine triplet	0.99 (0.88 to 1.11)	0.98 (0.84 to 1.14)
Doxil triplet	1.00 (0.89 to 1.12)	0.97 (0.83 to 1.14)
Topotecan doublet	1.09 (0.98 to 1.22)	0.07 (0.92 to 1.24)
Gemcitabine doublet	1.05 (0.94 to 1.18)	1.04 (0.88 to 1.21)

* CI = confidence interval; GOG = Gynecologic Oncology Group; ICON = International Collaborative Ovarian Neoplasm Studies.

primary outcome measure at the three stages are safety at the first stage, progression-free survival at the second stage, and overall survival at the third stage.

Discussion

We have proposed the MAMS design as a direct strategic response to the pressing need for clinical trials to achieve more reliable results more quickly. Key to the use of the design is the principle that many potential new therapies need to be tested in similar time frames. The MAMS design may be an appropriate alternative to the traditional phase 2 followed by phase 3 trial setting in certain situations (Box 1). Our approach has two distinguishing characteristics: we compare many new therapies at once against a control treatment and reject insufficiently active therapies on the basis of an intermediate outcome measure in a randomized pairwise comparison with the control. This "unified" approach gains its speed from the fact that many therapies are considered at the same time and that there is a planned and seamless move from one stage to the next. The reliability of this design stems from the use of an appropriately powered randomized comparison on an intermediate outcome measure. The GOG-182/ICON5 trial clearly shows the practical value of MAMS trials. With three real examples, we hope that we have shown that such trials are feasible and can lead to major improvements in speed and decision making.

Multi-arm, single-stage trials are not new—many such trials have been performed in different diseases in different parts of the world (27–29). Although such trials might initially appear complex particularly to patients, clinicians, competent authorities, ethics committees, and trial oversight committees, concerns about the feasibility of recruitment to such trials have not been realized. In STAMPEDE, a two-part patient information sheet was used to aid in the understanding of the design. Patients were provided with a summary of the trial and its arms at the beginning and were given more detailed information about their particular arm after random assignment. This "two-stage" informed consent process is in the process of being adopted more widely for more conventional trials by ethics committees in the United Kingdom.

Box 1. Summary of when a multi-arm, multi-stage (MAMS) trial may be useful.

A MAMS design may be useful when:

1) Many new approaches (therapies/regimens) are available for evaluation in phase 2/3 trials:
   i) that have sufficient promise to warrant investigation
   ii) that can be distributed widely
2) There is no a priori reason to expect one approach to be better than another
3) There is an intermediate outcome measure that is correlated with the primary outcome measure (the primary outcome may serve as an intermediate outcome measure if it can be measured at several time points) such that:
   i) If there is little or no impact of an experimental arm on the intermediate outcome, there is likely to be little or no impact on the primary outcome.
   ii) If the intermediate and primary outcome measures are measured on the same scale, then, if the alternative hypothesis is true on the primary outcome, the alternative hypothesis (or something more extreme) is also likely to be true for the intermediate outcome measure.
4) There are sufficient funds to support a more complex MAMS trial
5) The accrual rate can support an MAMS trial

The multi-stage component adds a number of staging posts at which accrual to each of the experimental arms can potentially be stopped when there is good evidence that the experimental arm is unlikely to be clinically better than the control arm. In other areas, this may be understood as a stopping guideline for "futility." Again, this is not a new principle, except that we propose using an intermediate outcome measure that allows us to screen out ineffective therapies. In situations for which an intermediate outcome measure may not be available, it may be possible to use the primary outcome measure measured earlier in time. Such an alternative approach is similar to the approach proposed by Simon et al. (30).

The intermediate outcome measure does not need to be a surrogate for the primary outcome measure. It does need to be related in the sense that if a new treatment has little or no effect on the intermediate outcome measure, then it will likely have little or no effect on the primary outcome measure. Importantly, however, this relationship does not have to apply in the other direction. Thus, we do not assume that just because an effect has been observed on the intermediate outcome measure that we shall see an effect on the primary outcome measure. Good examples of intermediate and final outcomes are progression-free survival and overall survival, respectively.

From one point of view, the early stages of the design (at which the intermediate outcome measure is being used) could be viewed as a set of simultaneous well-designed comparative randomized phase 2 trials. The main difference is that there is a formal randomized comparison that is appropriately powered and designed to inform stop/go decisions, in contrast to the traditional nonrandomized comparisons that are made in the conventional testing of new therapies. At these early stages, the probability of continuing to the next stage should the alternative hypothesis be true should remain high—we have typically used 95%. To achieve this high probability, the probability of continuing to the next stage should the null hypothesis be true is relaxed—we have used 10% to 50%. This probability can get progressively smaller as the information (number of events) increases, the STAMPEDE trial is a good example. The type I error over the trial is protected by the fact of the need to jump each staged hurdle. The likelihood of an ineffective therapy passing through all intermediate stages and the final stage is small indeed. This component of the design can be considered to provide a seamless transition from phase 2 (earlier stages of the trial) to phase 3 (final stage), with all patients involved in the earlier stages contributing to the final stage, and as such has similarities to other seamless phase 2/3 designs. A review of these types of designs and their application has been provided by Schmidli et al. (31).

The two components of the MAMS design can be used separately. For example, a staged design could be used in a two-arm trial, which would be more efficient than a traditional two-arm trial. Thus, the trial design would allow for early stopping for futility. Alternatively, a multi-arm trial could be performed with only one stage. Although both may give some benefits, they do not reap the full benefits of the MAMS trial.

This MAMS design also forces those who are designing new trials to think more strategically beyond the question of "We have a promising new compound, can it improve outcomes for patients with disease x?" to "How can we plan to improve outcomes for patients with disease x as swiftly and reliably as possible?" The first question would perhaps lead to a traditional large-scale two-arm phase 3 trial, whereas the latter should lead to widespread considerations of the different experimental arms available at any given time. As such, this design may be particularly pertinent to researchers and agencies in the public sector. There are also advantages in the flexibility allowed by such designs. For example, different arms do not necessarily need to include different agents; they could explore different durations or doses of a new agent, such as in ICON6—a three-arm, three-stage trial of the new targeted agent cediranib. This approach may be particularly important for such targeted agents for which the optimal duration or dose of therapy is often unclear when initiating phase 3 trials.

The MAMS design is not without potential drawbacks. Although the trial itself may be of shorter duration, it may take longer to set up. A contributing factor is the greater deal of complexity that arises when drugs may have to be sourced from different companies. Perhaps surprisingly, industry partners have been supportive of this design, even when it puts their agent into the same trial as a competitor's. Strategies that we have used to persuade industry partners to include their agents in such trials are as follows. First, it would be to the company's detriment if their product was not included and a promising agent from another company may take its place. Second, the MAMS design does not compare "head-to-head" the various products from different companies; each experimental arm is compared formally only against the control arm, and this is not dissimilar to running separate two-arm trials. It is possible that all of the experimental therapies will prove successful, and there may be an opportunity to further improve outcomes by looking at combinations of the experimental therapies. Finally, the design is a form of risk management for the company (and the investigators). If an experimental therapy is unlikely to prove beneficial, then it is better to stop investing further patients, time, and money in testing it. The three examples show that these approaches have been successful with a wide range of drugs from a number of companies.

In certain situations—for example, when more than one of the experimental arms continues to the final stage—then MAMS trials will need to be considerably larger than standard two-arm trials. Such large trials will often require cooperative groups to undertake them and further may require international collaborations. Despite aims to improve the harmonization of the regulatory environment (32,33), international collaborations are complex to initiate and undertake. It is also unlikely that individual pharmaceutical companies will have more than one product that they are willing to test in a particular setting at any given time. All of these issues mean that MAMS trials are likely to be possible only in cooperative groups. These groups do, however, undertake a large proportion of the large-scale cancer trials. Software that is available from Stata has been developed to help design MAMS trials and may be obtained from the authors upon request.

Our hope is that others will exploit the opportunities that the MAMS trial design offers to correspondingly speed up the assessment and introduce new therapies to patients with a wide range of cancers, and also more broadly in other diseases.

References

1. Bryant PA, Venter D, Robins-Browne R, Curtis N. Chips with everything: DNA microarrays in infectious diseases. *Lancet Infect Dis.* 2004; 4(2):100–111.
2. Lane D. The promise of molecular oncology. *Lancet.* 1998;351(Suppl 2): SII17–SII20.
3. Parexel. *Parexel's Pharmaceutical R&D Statistical Sourcebook Parexel International;* 2006.
4. The Royal Society. *Personalised Medicines: Hopes and Realities.* 2005;http://www.royalsoc.ac.uk/document.asp?id = 3780. Accessed 21 July 2008.
5. Food and Drug Administration. *Innovation or Stagnation. White Paper.* Washington, DC: Food and Drug Administration; 2004.
6. Webster B, Lewison G, Rowlands I. *Mapping the Landscape II: Biomedical Research in the UK, 1989–2002.* London: City University; 2003.
7. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ.* 2003;22(2):151–185.
8. EORTC. Phase II trials in the EORTC. *Eur J Cancer.* 1997;33(9): 1361–1363.
9. Roberts TG, Lynch TJ Jr, Chabner BA. The phase III trial in the era of targeted therapy: unraveling the "Go or No Go" Decision. *J Clin Oncol.* 2003;21(19):3683–3695.
10. Bailar JC, Gornik HL. Cancer undefeated. *N Engl J Med.* 1997;336(22): 1569–1574.
11. Saltz LB, Clarke S, Diaz-Rubio E, et al. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. *J Clin Oncol.* 2008;26(12):2013–2019.
12. Vogelzang NJ, Rusthoven JJ, Symanowski J, et al. Phase III study of pemetrexed in combination with cisplatin versus cisplatin alone in patients with malignant pleural mesothelioma. *J Clin Oncol.* 2003;21(14):2636–2644.
13. Kumar A, Soares H, Wells R, et al. Are experimental treatments for cancer in children superior to established treatments? Observational study of randomised controlled trials by the Children's Oncology Group. *BMJ.* 2005;331(7528):1295–1300.
14. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov.* 2004;3(8):711–715.
15. Giaccone G, Herbst RS, Manegold C, et al. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small-cell lung cancer: a phase III trial—INTACT 1. *J Clin Oncol.* 2004;22(5):777–784.
16. Herbst RS, Giaccone G, Schiller JH, et al. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial—INTACT 2. *J Clin Oncol.* 2004;22(5):785–794.
17. Kris MG, Natale RB, Herbst RS, et al. Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non-small cell lung cancer: a randomized trial. *JAMA.* 2003; 290(16):2149–2158.
18. Fukuoka M, Yano S, Giaconne G, et al. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small cell lung cancer. *J Clin Oncol.* 2003;21(12):2237–2246.
19. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics.* 1998;54:1014–1029.
20. Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes, with an application in ovarian cancer. *Stat Med.* 2003;22(14):2239–2256.
21. Stampede Trial Development Group. *Stampede Protocol.* London: MRC Clinical Trials Unit; 2004.
22. Bookman MA; the Gynecologic Cancer InterGroup (GCIG). GOG0182-ICON5: 5-arm phase III randomized trial of paclitaxel (P) and carboplatin (C) vs combinations with Gemcitabine (G), PEG-liposomal doxorubicin (D), or topotecan (T) in patients (pts) with advanced-stage epithelial ovarian (EOC) or primary peritoneal (PPC) carcinoma. *Proc ASCO.* 2006. Abstract 5002.
23. ICON6 Trial Development group. *ICON6 Protocol.* London: MRC Clinical Trials Unit; 2007.

24. Bookman MA, Malmstrom H, Bolis G, et al. Topotecan for the treatment of advanced epithelial ovarian cancer: an open-label phase II study in patients treated after prior chemotherapy that contained cisplatin or carboplatin and paclitaxel. *J Clin Oncol.* 1998;16(1):3345–3352.
25. O'Reilly S, Fleming GF, Baker SD, et al. Phase I trial and pharmacologic trial of sequences of paclitaxel and topotecan in previously treated ovarian epithelial malignancies: a Gynecologic Oncology Group study. *J Clin Oncol.* 1997;15(4):177–186.
26. Lyass O, Uziely B, Ben-Yosef R, et al. Correlation of toxicity with pharmacokinetics of pegylated liposomal doxorubicin (Doxil) in metastatic breast carcinoma. *Cancer.* 2000;89(5):1037–1047.
27. Kleeberg UR, Brocker EB, Lejeune F, et al. Adjuvant trial in melanoma patients comparing rIFN-a to rIFNy to Iscador to a control group after curative resection of high risk primary (≥3mm) or regional lymph node metastases. *Eur J Cancer.* 1999;35(1):582 (abstract 24).
28. PENTA. Comparison of dual nucleoside-analogue reverse-transcriptase inhibitor regimens with and without nelfinavir in children with HIV-1 who have not been previously treated: the PENTA 5 randomised trial. *Lancet.* 2002;359(8532):733–740.
29. Mutabingwa TK, Anthony D, Heller A, et al. Amodiaquine alone, amodiaquine+sulfadoxine-pyrimethamine, amodiaquine+artesunate, and artemether-lumefantrine for outpatient treatment of malaria in Tanzanian children: a four-arm randomised effectiveness trial. *Lancet.* 2005;365(9469):1474–1480.
30. Simon R, Thall PF, Ellenberg SS. New designs for the selection of treatments to be tested in randomized clinical trials. *Stat Med.* 1994;13(5–7):417–429.
31. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biom J.* 2006;48(4):635–643.
32. European Medicines Agency. ICH Topic E6(R1), guideline for good clinical practice. http://www.emea.europa.eu/pdfs/human/ich/01359en.pdf. Accessed 21 July 2008.
33. European Commission Enterprise and Industry, The Rules governing medicinal products in the European Union. http://eceuropa.eu/enterprise/pharmaceuticals/eudralex_en.htm. Accessed 21 July 2008.

Funding

The MRC Clinical Trials Unit has received educational grants and free or discounted drugs from some of the companies involved in the MAMS trials described in this commentary. MAB is a member of the Data Monitoring Committee Genentech Oncology (compensated) and is on ad hoc advisory boards and/or received lecture honoraria from Bristol Meyers Squibb, Glaxo SmithKline, Eli Lilly, Genentech Oncology, Novartis, Sanofi Aventis, and Johnson & Johnson.

Notes

The sponsors had no role in the preparation of the manuscript or the decision to submit the manuscript for publication.

Manuscript received July 7, 2007; revised June 6, 2008; accepted July 3, 2008.