Novel designs for clinical trials in multiple sclerosis: another excursion to Sirenum scopuli?

Stacey S Cofield[1†] & Gary R Cutter[2]
[1]University of Alabama at Birmingham, Department of Biostatistics, 410C Ryals Public Health Building, 1665 University Boulevard Birmingham, AL 35294-0022, USA
Tel. : +1 205 934 4932; Fax: +1 205 975 2541; E-mail: scofield@uab.edu
[2]University of Alabama at Birmingham, Department of Biostatistics, 410b 1665 University Boulevard, Birmingham, AL 35294-0022, USA
cutterg@prodigy.net

This review summarizes a number of key approaches to novel designs in multiple sclerosis clinical trials. The concept of the novel design is discussed, and the interplay between design decisions and trial end points is debated. We provide reminders that there are trade-offs between the end points chosen and the type of design. The opportunity for more novel designs resides in Phase II studies, and several types of novel designs are discussed, including: adaptive designs, Bayesian designs, treatment strategy designs and futility designs. The fact is that the end points in multiple sclerosis still exert a limiting effect on the designs, and underscore the need for long-term follow-up information compared with short-term information. Overall, the design is only as good as the underlying question and the end points chosen to evaluate it.

With the advances in treatments for multiple sclerosis (MS), a host of considerations are occurring in the MS clinical trial community. There has been the realization that the road ahead is going to be littered with agents that have no improved efficacy, coupled with new agents that may not be better in effectiveness, but will be safer or more convenient; or those that may be better in efficacy with unknown long-term safety. The search will also continue with a range of considerations for treatments in stages of the disease for which therapies to date have been disappointing. Effective treatments and a rising standard of care require larger sample sizes to detect meaningful differences. This has already occurred in clinical trials of this decade. In the 1980s, trials often involved fewer than 100–200 patients. In the 1990s, the numbers increased to 300–600, and currently the sample sizes are routinely approaching 1000 patients per trial, and even Phase II trials can number in the hundreds. The magnitude and cost of such trials has led to an interest in novel study designs that can either provide faster answers to a host of questions, or answers to questions utilizing fewer study subjects. But will these new designs result in useful and meaningful results for patients? Like the sailors in Greek Mythology that were lured towards the rocky islands of Sirenum scopuli by the beautiful songs of the Sirens only to meet an untimely end, is the quest for novel designs merely another excursion to Sirenum scopuli? Are novel designs the Sirens, calling researchers and clinicians, only to have them crash into the perilous rocks of study futility?

Every grant submitted to the NIH is evaluated on its level of innovation, so it is surprising that this Siren's call for novel designs in MS clinical trials is both enchanting and almost unavoidable. An interesting comment on clinical trial design was put forth by Von Hoff some 10 years ago [1], where he pointed out that "there are no bad anticancer drugs, only bad clinical trial designs". This interesting lecture lamented the fact that 90% of the anticancer agents developed in the laboratory never made it to routine clinical use. His arguments are that "*(a) the toxicities of the agent were too great; (b) there was a lack of efficacy; and (c) no attention was paid to the mechanism of action of the compound when the clinical trials were designed and conducted."* The latter of these three reasons clearly challenges designers of trials to come up with, or at least consider, novel designs in this era of targeted molecules.

A discussion of novel designs must first begin with a definition of a novel design. Novel is defined as 'new', 'fresh' or 'original', and of course design is defined as 'to formulate a plan for', 'devise', 'structure' and so on. This comes as no surprise, but in the context of MS, does this mean new to MS, or a newly invented design first appearing in MS? For this discussion, novel approaches will be focused on the following aspects that are essential to all clinical trails: study phase, outcome selection, analysis methods and sample size selection. Determining any one of these requires defining the others. That is, to determine the sample size necessary for the study, one must first define the question and the outcome to determine the analysis method that will be used for sample size estimation. For the

Keywords: adaptive designs, Bayesian designs, clinical trials, hypotheses, novel designs, Phase II, Phase III

future medicine part of fsg

purposes of this discussion, 'novel design' is defined as an approach in one of these areas that has not been used frequently in MS research, and thus could be considered unique or novel in this field.

Study phases

The very first task any investigator must tackle is defining the question(s) to be answered. While this may seem a simple task, clearly defining the question is often the most difficult step in designing a clinical trial. Defining the question is often intertwined with the phase of the trial that will be used. Clinical trials are classified into four Phases (I–IV):

- Phase I trials focus on dose finding and schedules of administration in animals and/or healthy humans based on pharmacokinetic and pharmacodynamic properties and characteristics, as well as the immediate safety of the treatment agent;
- Phase II trials are utilized to evaluate, in appropriate patients, preliminary safety and treatment efficacy of the agent under study, and are often called proof-of-concept trials;
- Phase III trials are utilized to demonstrate both overall safety and effectiveness; often called pivotal trials because they change the practice of medicine;
- Phase IV trials are conducted after the drug or treatment has been approved to obtain information on the drug's effect in other populations and determine side effects with long-term use; often called postmarketing studies.

Phase I trials are often implicitly based on the assumption that increasing the amount of agent increases the corresponding response, and that there is some direct generalization from animals to humans. Phase IV studies are just being initiated in MS, and while a number of novel design considerations have occurred in both Phase I and IV designs in many diseases, they are not the focus in this paper. We will concentrate our discussion on Phase II and Phase III designs. Unlike Phase I studies, Phase II and III studies are conducted in participants with MS, and unlike Phase IV, the treatments are tested in rigid clinical settings, usually involving controlled comparisons.

Phase II & III studies

These phases can be based on multistage designs, where the first Phase II trial might look at dose finding in MS patients, and a second study be aimed at estimation of effectiveness or specific safety issues. While multiple studies are often necessary, a single study can lead to the next key phase, Phase III trials. Phase III trials are generally directed at standard clinically meaningful end points, can evaluate potential new end points, and are usually large multicenter and perhaps even international trials.

Phase II trials offer a far greater range for novelty than Phase III trials. A simple reason is that Phase II trials have multiple goals and even an exploratory nature compared with Phase III, where there is an almost central focus on specific effectiveness (the primary end point) and safety. Phase II trials seek to define the best dosage choice to achieve efficacy, the presumed safest dose, the responsiveness of various end points, the appropriate target populations and so on. In Phase II we are seeking the clearest answer in the shortest time with sufficient evidence (proof-of-concept) that will enable us to design an efficient Phase III trial.

Regardless of the trial phase being implemented, clinical trial designs depend first and foremost on the question being asked and the end points chosen to assess the hypothesis. This is the fundamental issue of any trial. There are three general classes of questions:

- Superiority trials
- Noninferiority trials
- Equivalence trials

Superiority trials aim to demonstrate that one treatment is better than another (**Box 1**), noninferiority trials attempt to show that one treatment is not worse than another and equivalence trials attempt to show a treatment is neither worse nor better than a standard treatment with tightly controlled equivalence limits. Most randomized clinical trials want to show that one treatment is superior to another treatment. In terms of MS, especially relapsing–remitting MS (RRMS), where there are a number of effective treatments, superiority may not be the hypothesis of interest. While noninferiority trials have not been commonly used in Phase III trials in MS, they can be used to show that a new less expensive drug, a drug at a different dose, or a drug with fewer side effects, is no worse than a standard treatment. Regardless of the hypothesis being tested, study designs often have multiple iterations related to the primary end point until a final design is chosen. In addition, defining the outcome of the study is integral in choosing the hypothesis, so that the minimum clinically significant difference can be chosen. That is, a difference in outcome that would be seen as important to a patient.

Box 1. Superiority, noninferiority, equivalence hypotheses.

Let the new treatment be T and the standard treatment be S. Also, let the minimum clinically meaningful difference be d; that is, let d be the largest difference that can be judged as being acceptable (and generally smaller than the difference observed in standard versus placebo trials). Given T, S and d, the three types of null (H0) and alternative (HA) hypotheses can be defined as:

- Superiority: H0: $T \leq S$; HA: $T > S$. The null is that T is worse than or the same as S, and the alternative is that T is better than S.
- Noninferiority: H0: $T \leq S - d$; HA: $T > S - d$. The null is that T is worse than or the same as S reduced by d, and the alternative is that T is better than S reduced by d.
- Equivalence: H0: $T \neq |S - d|$, HA: $T = S - d$. The null is that T is different (better or worse) than S reduced by d, and the alternative is that T is the same as S reduced by d, where d can be zero.

Note that a superiority test is a noninferiority test where d = 0; that is, there is no acceptable minimum clinically significant difference.

Outcome selection

Obviously, within each class of trial, safety is a major concern. There are intellectually (and some times formally) dual considerations: efficacy and safety. What is an acceptable level of efficacy given safety concerns, or what are acceptable safety risks given the high levels of efficacy?

In MS we have numerous outcome measures and no gold standard. The Expanded Disability Systems Scale (EDSS) is frequently utilized because of its familiarity; however, given that the EDSS is not a continuous scale, the use of average changes can be misleading when applying clinical trial results to patient expectations. The Multiple Sclerosis Functional Composite (MSFC) was developed to provide a paradigm shift enabling MS researchers to consider a measure with some enhanced mathematical properties. However, it lacks easy clinical interpretation and, like the EDSS, suffers some performance issues in particular patient populations. Relapse rates are a common outcome in MS trials of RRMS, but these too are only correlates of disease progression and have a number of problems of measurement. Relapses are often confirmed by changes in the EDSS, and even when utilized as a component of relapse rates, the EDSS can vary widely depending upon the population being studied and who is measuring the subject. The MRI is a highly technical method of assessment with quantifiable characteristics, but again the clinical interpretability has yet to be definitively shown. In addition, which measure or combination of measures (gadolinium-enhancing lesions, T1 or T2-weighted counts or volumes, atrophy measures and so on) are best for each type and stage of MS has yet to be determined. Regardless of the outcome chosen, relapse rates, the EDSS, MSFC and MRI measures are certainly not as clear as mortality as an outcome in cardiovascular or oncology trials [2]. This can present a problem with novel designs that often rely on definitive and clearly defined end points.

Each of the MS outcome measures has also been characterized in a number of different ways. For example, change in EDSS has been used as a key outcome measure in earlier studies, but sustained change in EDSS over 3 or 6 months has been used to establish a true change from measurement error or transient worsening in more recent trials. In addition, the attainment of benchmark levels of EDSS has been utilized, for example, the percentage of participants or time to reach an EDSS score of 4 or 6. The use of the same measure in various ways can be considered novel in the design of a MS trial because each modification is designed to provide key information in a way appropriate to the question being addressed. In many other diseases a great deal of effort has been placed on defining uniform outcome measures. In cancer, there are standardized definitions of event-free survival, progression-free survival, disease-free survival, relapse rate, death, complete response, time to progression and so on. However, from the length of this list, one can see that no single outcome or event is sufficient, even in the face of harder end points, and despite greater specificity in outcome definitions. Thus, the lack of clearly defined end points, while a drawback in MS, is certainly not a major deterrent to novel designs, and it should be clearly stated that MS is not unique.

Analysis methods & sample size approximation

Determining the sample size necessary for a trial not only requires the definition of the outcome, but also requires selecting an appropriate analysis method. The chosen analysis method will ultimately lead to estimation of the trial sample size. A good deal for novel innovation has been carried out in the arena of sample size estimation. Sormani *et al.* devised more accurate sample size estimates based on averaging multiple MRIs to minimize misclassification of active and inactive subjects [3,4]. They achieved this by simultaneously deriving sample size estimates that were based on statistical models that more appropriately captured the increased variability often found in trials, as opposed to simplifying assumptions to estimate sample sizes, which is the more common method utilized.

Adaptive designs are another widely discussed innovation of clinical trials. The concept is considered novel, and it should be noted that while popular in discussions, few, if any, have been accepted to date by the US FDA in Phase III trials. The problem is not with the novelty of the design, but with the decision rules and the implementation of the changes required to make the adaptations. Adaptive designs move theoretically in a seamless manner between an early development phase (Phase I and II) and a registration phase (Phase III). They gain the same information as separate Phase trials, but do not have the usual gap in time between Phase II and Phase III, which often runs between 1 and 2 years. The ideas are straightforward:

- Observe a portion of the data being generated
- Check the assumptions of the design
- Adjust the design, usually via sample size or study duration, and continue with the trial

This process would shorten the development time by eliminating noninformative or poor-performing doses of drugs and/or cutting out the lag-time between Phase II completion and Phase III enrolment. So why wouldn't one always use these methods?

One reason is that the sponsor is committed in advance to the corrections and end points of the Phase II trial. This places a lot of planning on the end points, restricts the use of additional knowledge gained on other secondary or exploratory end points during the trial and requires the decision rules to be carefully and clearly spelled out in advance. The decision to add patients indicates that something is wrong in the initial assumptions, and depending on how the adaptation is to work, could diminish the enthusiasm of physicians, participants and even the sponsor for the increased sample size and/or its cost. Nevertheless, altering the design when the assumptions are incorrect, such as a vastly overestimated benefit or a much lower rate of events than what was expected, while maintaining sufficient power, requires a substantial increase in the sample size. Such information may not inspire investigators to continue enrolling subjects. Other logistical issues involve when to make these adjustments, and who has access to the underlying data to make the determinations. A problem for MS is that event and progression rates do not tend to be constant in time and, thus, if the trial adjustments are made too soon, they might not be representative of the entire study period. Conversely, waiting too long to adapt the trial may result in increasing the trial size near the time when enrolment is almost complete, or even re-opening enrolment that was completed. Using a classical Phase II approach followed by a Phase III enables the same adjustments and others, but with the standard administrative lag to get sites and institutional review boards on side with the Phase III trial, a process that can take 1–2 years or even longer if the initial analyses require internal acceptance before continuing.

A bigger Siren, as the Greek Mythology calls them, is that of Bayesian methods. There are two major philosophies of statistical analyses: one is the classical or frequentist approach to trials, and the other is the Bayesian approach. Bayesian methods are gaining a lot of support and discussion about their value to the interpretability of results and the utility of the methods. However, as with all Sirens, it is best to understand the waters surrounding the call.

The frequentist school of thought asks questions of how likely is this result if one were to repeat this experiment over and over again, assuming what was originally hypothesized to be true? The Bayesian approach is interested in how likely the true state of the world is, given what has been observed now and in the past. Clearly both approaches have merit and a host of assumptions. Violating the assumptions in either case can lead to problems. Historically frequentist methods have been most commonly used in clinical trials, and familiarity breeds contempt. Thus, the beauty of Bayesian methods seems to be a novel solution for not only MS trials, but other disciplines as well.

The Bayesian approach uses information in a way that enables one to use prior information to make an informed assessment and then updates that assessment with each new set of data. Frequentist approaches asks questions of an idealized situation with an eye as to whether that which is observed is likely to have arisen from the underlying hypothesized situation. Nevertheless, in terms of clinical trials, one might argue that beyond the philosophical approach to interpreting the results of a trial, the idea of incorporating prior information into the final analysis is a fundamental difference between the approaches. The fundamental question is whether you believe a clinical trial is designed to find the best estimate of the treatment effect, or whether you are desirous of an independent demonstration of the treatment effect. It is this perspective that leads to greatly different views

Box 2. Bayes Theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\text{Not A})P(\text{Not A})} = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ is the probability of event A occurring given event B has already occurred, where $P(A)$ is the probability of event A occurring.

on the value of Bayesian statistics in Phase III trials, whereas there is much less controversy over their use in Phase II trials where feasibility and estimation are paramount interests.

Generally, Bayesian analyses use information that is available from the past to assess the situation of the groups, and then modifies that information as new information is obtained. The fundamental approach for this form of analysis is a mathematical relationship developed by the Reverend Thomas Bayes (1702–1761), which allows one to infer the probability of one event given some condition from a series of related, but not exactly similar probabilities. Put simply, Bayesian methods can be used to estimate the probability of an event A happening given an event B has happened, using a series of previously observed probabilities: 1) the probability of A occurring (with no conditions), 2) the probability of A not happening (with no conditions), 3) the probability of B, given A has occurred (one condition, that A has occurred) and 4) the probability of B occurring given A has not occurred (**Box 2**). Even more simply, the probability of A happening given B has happened can be determined with 1), 3) and the probability of B occurring (no conditions). While this may seem a bit confusing, put in the context of a patient scenario, one might like to know what is the probability that in 2 years a patient would have a change in their EDSS of 2 or more (event A), given the patient has disease duration of less than 8 years (event B)? In a prospective clinical trial, you would gather patients with less than 8 years of disease duration, observe them all for 2 years and measure their EDSS changes, and then calculate the observed probability. However, using Bayes theorem, if you have the other related probabilities, you can estimate the probability of interest without following any new patients, simply by invoking the equation developed by Bayes.

In the placebo groups of the trials collected to develop the MSFC [6], there were 986 subjects that had both a baseline and 2-year EDSS values as well as disease duration. A total of 529 (529/986; 53.7%) of these patients had a disease duration of less than 8 years at their time of entry into the trial database. So, with this data, what is the probability that a patient will progress by 2 or more EDSS units over 2 years, given that they have been diagnosed with MS within the past 8 years? The statistical formula for this is presented in **Box 3**.

Knowing that 53.6% of subjects had a disease duration of less than 8 years, the resulting conditional probability can be determined from the probability of having a duration of less than 8 years given a change of 2 (108/175; 61.7%) and the probability of a change of 2 (175/986; 17.7%). Thus, the overall updated probability of progression given that you started the study with a disease duration of under 8 years would be:

$$\{0.617 \times 0.177\}/0.536 = 0.203 \text{ or } 20.3\%$$

The naive estimate of 17.7% progression, regardless of duration of disease, can thus be refined or updated to 20.3% given the knowledge of disease duration at baseline. This is how Bayesian analysis works.

This use of prior information would tell us that a patient who is within 8 years of diagnosis is more likely to have a change of 2 points on their EDSS compared to simply estimating the probability of changing by 2 points. We could then embark on a new study using these same principles, and modify these probabilities of progression based on the new information as it is obtained. Bayesian statistical estimates of treatment differences work similarly, but are based on estimating these population parameters to assess whether the progression rates of two groups differ. Reapplying this formula to data and using the updated probabilities to gain insight into how a treatment is working is used to analyze the results of a Bayesian trial.

While Bayesian statistical methods offer more than this simple probability estimates, the use of prior data to update future data through this theorem is the essence of the process. Our revised or

Box 3. Application of Bayes theorem.

$$P(\geq \Delta 2|<8) = \frac{P(<8|\geq \Delta 2)P(\geq \Delta 2)}{P(<8|\geq \Delta 2)P(\geq \Delta 2) + P(<8|<\Delta 2)P(<\Delta 2)} = \frac{P(<8|\geq \Delta 2)P(\geq \Delta 2)}{P(<8)}$$

updated estimate is informed by our prior information. In Phase II studies, use of all available information at the start of a trial to update our best estimates of the outcomes of interest may indeed be one of the best uses of Bayesian methods. These novel methods may formally combine past information to better enable careful planning of Phase III trials.

There are two major assumptions of these procedures that must be kept in mind. First, the process of setting up a Bayesian study process often requires much more time and prior information. The mathematical relationship of the past distribution (the prior) must be integrally tied to the updated distribution (the posterior). Importantly for MS, which does differ dramatically from other diseases such as cancer and heart disease, there is an implicit assumption that the natural history or treated history of the disease has not changed. Such an assumption is perilously dangerous in MS, where the definition of the disease itself has changed and been modified in the past decade, and the trial populations available to model the Bayesian designs have also changed dramatically.

These latter changes are in part due to clinicians possibly treating particularly high-risk patients rather than entering them into trials. In addition, the changes may be due to the global shift to utilizing new sources of patients, such as those from Eastern Europe who may or may not behave in a similar fashion to existing trial data. In addition, the recent changes in disease definitions, particularly in RRMS, has allowed for earlier diagnosis and treatment of patients, ultimately altering the landscape of 'historical populations'.

Bayesians would argue that prior data are not essential to using these methods in practice, because you can assume a so-called noninformative prior. This is true, but when such assumptions are invoked, the Bayesian methods may require much larger sample sizes in which to achieve the precision desired. Bayesians would still argue that their methods are more intuitive than frequentist methods, and should be used for their philosophical advantages.

Another major advantage of Bayesian designs is that problems of multiple comparisons in effect disappear, since the approach is merely geared to summarizing the outcomes rather than providing a p-value. This affords continuous monitoring without concerns for how much testing is carried out. However, this advantage can also be a major disadvantage. When Bayesian methods are used, greater *a priori* decision-making and rules of operation are required. Even the decision to stop must be described in advance and the characteristics of how the models will operate must be specified. Such specification requires detailed critical thinking in advance, which requires a consensus opinion that is often difficult to obtain amongst the various parties in the planning of a trial. Thus, while the novelty of Bayesian designs holds great promise for being intuitively understood, the reality of their use in MS clinical trials may not be seen in the immediate future.

If we cannot make use of the Sirens of Bayes, are there other opportunities in Phase II to move forward? Certainly, the temptation to trade the speed of obtaining an answer (time) with increased sample size versus longer follow-up is still present. As noted above, the mechanism of action may be critically tied to the design considerations. Designs that focus on the mechanism of action rather than a clinical outcome may become even more important as MS moves from drugs that block or eliminate inflammation (the patient feels better but continues to progress in disability) to drugs that repair or protect neurons (halt progression or even repair the deficits).

Designing trials that focus on neuroprotection and neuro-repair must critically face the potential timing of the effects. For example, a drug purported to have solely neuroprotective effects may require longer term studies to allow the early failure due to axonal damage already initiated. Such situations exist in many cardiac and cerebrovascular surgeries, where the first 30 days are treated separately from longer term survival outcomes due to expected increases in morbidity and mortality from the surgeries themselves.

A concept that is not new in a statistical sense, but is novel to MS, is the use of repeated measures. With the use of annualized relapse rates (ARR), only a single measure is used per patient (number of relapses per years on study). The use of the ARR, or any other single measure such as time to first relapse, or relapse free, ignores the pattern of events within subjects over time. It may be of interest to determine not only if the groups are different, but how they are different: do events occur evenly over time? Is there a remission period followed by an increased event period? Or is the treatment not immediately effective, but over time becomes more effective? Using repeated measures on the same person to assess relapses may be shown to be a more appropriate method of analysis, but this can increase estimated variance and will not lead to smaller sample sizes [5]. In addition, while novel in MS,

given the long length of the disease course relative to the often short (less than 3 years) duration of clinical trials, relying on a repeated measures analysis may not have any advantages in terms of short-term efficacy. This may change if patients can be followed for longer periods of time in controlled situations.

Treatment strategy designs

Intertwined with the issues associated with the phase of the study, the outcome being analyzed and the methods and sample size to be used, how the patients will be treated can also offer some novel approaches. MS treatments have basically been focused on a 'one-size fits all' approach to therapy. Only recently have formal combination therapy trials been undertaken. However, the experience with natalizumab in combination with interferon β-1a (Avonex®), suggesting that the pair lowered the immune system to a potentially unsafe level, may be a harbinger of future problems with multiple-treatment approaches. In many other diseases, stepped care regimens or strategies of therapy have been invoked to use increasing amounts of drug or combinations only on those patients who meet some failure or lack of success criteria. These designs would be novel in MS today, but are going to be the likely consequence of the increasing number of options for MS patients in the future. Designs that attempt to look at treatment strategies offer both a host of opportunities and some compromises in the usual purity of naive treatment groups. Designs that focus on induction therapy followed by maintenance can be considered in this class. At present, we still could conduct such a trial using a single agent or placebo followed by maintenance therapy in a classic two-group design. However, as MS treatment evolves, we may need to consider multiple paths of treatment. For example, we might start with an interferon, add glatiramer acetate if a number of relapses or progression occurs, and switch to natalizumab (or another agent) if the disease is not kept under control. As the number of pathways increase, the sample size grows as per the number of potential paths, but these treatment strategy designs may become more important in assessing the cost–effectiveness and/or cost:benefit of treatments.

Growth modulation index designs

When there is an increase in the numbers of patients previously exposed to a host of drug treatments, the number of options for treatment in trials is usually diminished by the exclusion criteria of the new trials. Under such situations, a number of potential patients who can be studied are eliminated. This situation was faced in cancer trials with a novel design for which the time to progression was the primary end point [7,8]. The design is predicated on the concept of a growth modulation index, a method suggested by Von Hoff in which each patient serves as his/her own historical control [1,9].

> *"The growth modulation index is defined as the ratio of a patient's time to progression on a Phase II cytostatic treatment, TTP2, relative to the time to progression observed from the patient's most recent prior anticancer treatment, TTP1, which serves as the patient-specific historical control value."* [1]

When the ratio exceeds 1.33, it is suggested that the new treatment has merit for future testing. This is roughly equivalent to a 30% increase in time to progression, a figure similarly used already in many MS trials. Such an approach might be useful in MS trials where time to sustained EDSS progression has occurred once, and watching until the next progression or stage is reached might be a plausible and ethical model for patients who need additional treatment. Slight modifications may be possible to develop novel study designs for MRI parameters such as T2 volume. Few protocols have been developed for 'failures' in MS, and even the definition of failure awaits a more formal explanation.

Random initiation designs & withdrawal designs

Random initiation and withdrawal designs are also useful for primary treatments and combination therapies where leaving patients untreated for too long a period is considered bothersome and/or unethical. These studies stagger groups of patients to be started on a drug at random times, and observe whether the ultimate therapy shows a time-dependent result. The continual questions about the best time to start therapy might be answered with a random initiation design coupled with a sufficiently long follow-up period to fully assess the benefits of early versus delayed therapy, where all patients are similarly followed in a masked fashion.

Withdrawal designs are also useful for assessing the question of how much drug for how long. In these studies, patients have one or more agents replaced by placebos at random times after study drug initiation and the long-term outcomes are tracked. These designs offer answers to such

questions as 'Does early elimination of inflammation suffice to control the patients with continuing treatment?' That is, can pulse therapy be used to control disease freeing an individual from continual exposure to the risks of drug therapy? While this has not been a question of high priority in MS treatment to date, future studies may need to address this concept if combination therapy is accepted, given the long-term costs of combination therapy or drugs with extremely long periods of action in the body.

Futility designs

Another impending problem in MS is the competition for patients to be studied in trials. As the number of potential treatments increase, competition increases for new patients or patients with minimal exposure to some drugs. This can slow the development of research by delaying recruitment or causing closure of studies, where interest in enrolment wanes, especially as other trials open for enrolment. One approach to screening more treatments in shorter periods of time is the futility design [10–12]. These designs are aimed at stopping trials that have a low chance of success, making continuation of such trials futile. Most Phase II and III trials focus on effectiveness of the treatment as the research question, also called the alternative hypothesis. The usual null hypothesis is that there is no difference between the treatment arms. In futility studies, one assumes that indeed the drug or treatment has benefit (null hypothesis), and the aim of the futility study is to reject this claim in favor of an inferior or no-difference outcome (alternative hypothesis). These trials can potentially be carried out using a single group in Phase II, with comparisons to historical controls shortening the duration and limiting the sample size.

In the single-group Phase II design, we might know from the standard of care and/or historical data the proportion of patients who respond or, in MS trials to date, mostly the proportion that fail. We are only interested in testing the new treatment if that new treatment improves on the standard failure rate by an amount Δ. In a standard design, we would hypothesize that the new and old treatments are the same, and devise a sample size designed to detect a difference of Δ or more. Suppose that we wanted to plan a two-group Phase II traditional study and feel that an absolute decrease of 10% in the failure rate (a 33% reduction) would be a significant improvement in outcome and worth pursuing for a Phase III development. A two-group χ^2 test of the proportion failing with a 0.05 two-sided significance level will have 80% power to detect the difference between the standard therapy Group 1 of 0.30 and the new treatment Group proportion of 0.200 (odds ratio of 0.583) when the sample size in each group is 294 subjects.

Such a study would be pretty large and limit the number of new treatments that could be tested. If we know from past data that the failure rate is approximately 30%, then we could design the trial as a futility study. Here we would hypothesize that the new treatment is better than the old, and our alternative is it is inferior to the old treatment. If we reject the hypothesis that the treatment is inferior, we would conclude it is futile to develop this drug. Here we are willing to assume that the 30% figure is correct, and that the new treatment is better than the old by 0.10. Now we use a single group to show that the difference between the groups is significant and, because we wish to screen drugs quickly and efficiently, we are willing to increase our type I error of calling a drug futile, then our Type I error might be set at 10%. In this situation, a sample size of 81 in a single group with a 0.10 one-sided significance level will have 80% power to detect the difference between the Null hypothesis proportion of 0.20 failure rate or better (0.3 is the standard) and the alternative proportion that it is no better than 0.30. Thus, with 81 patients we can determine if a drug is futile using this design, compared with nearly 600 patients for our typical two-group treatment A versus treatment B design.

The main benefit is of course not investing the time, money and effort in treatments that have low likelihood of return. Schwid and Cutter identified a number of disadvantages to these designs:

- The designs require a good knowledgebase from which to estimate the design parameters, and have limited ability to stop prior to observing a decent proportion of the planned Phase II trial
- Shortening the time and limiting the sample size may diminish the information on safety that is essential to moving a therapy forward in development
- Treatments with delayed effects may be missed, and for treatments of neuroprotection, this could be a major drawback
- Historical controls are weak controls, especially in an evolving disease and therapeutic era
- As the conclusions from these shortened, and smaller futility studies may not be sufficient to declare futility, they can increase the costs of

the Phase III endeavors because of the need for increased sample sizes to detect feasible but weaker treatments [11].

Limitations with MS

An important consideration, which is neither new nor novel, is the duration of the trials. As we move forward we should be forceful in examining short-term versus long-term effects. Pharmaceutical-sponsored Phase III studies are aiming to be as brief as necessary for registration, while the treatment consideration to the practitioners are to gain as much long-term information as possible. These competing interests are extremely costly and cannot be mere intellectual arguments, as studies may cost hundreds of millions of dollars during this phase. However, one recommendation might serve the scientific community and the patient populations well: follow all patients on their assigned treatment until the last patient completes the trial. This would result in more person-years of exposure being observed under controlled conditions, providing longer term results on which to judge the value of therapies. Possibly more important is that the added person-years of exposure will enable longer term assessment on specific side effects, overall safety and long-term consequences of therapies. Failure criteria may be added to ensure best medical care is applied to patients beyond the end of the formal trial period, but added information will benefit all parties at a fraction of the cost of future long-term studies that are attempted from registries. Long-term open-label studies or assessments of registries are at best limited, and while some information may be garnered by using propensity scores to adjust for many of the biases in assigning treatments to patients, these studies will fall short of clear and convincing outcomes. This is especially true in the face of marketing strategies designed to credit one treatment, while discrediting another.

As with any design, novel designs depend on quality end points. The MS community has moved forward in multiple domains on defining and evaluation of better end points. The linkage amongst these end points and the need for long-term rather than short-term results may lead to more novelty in design, and certainly more focused understanding of the results of trials, novel or otherwise.

Future perspective

There are many novel approaches available from within the MS research community and other disease models. The growth of treatments for MS has broadened the pool of researchers, and with that broadening cross-fertilization from multiple disease experiences. Enhancing the collaboration of experienced people from a variety of diseases will continue the cross-fertilization of design ideas. The reality of novel designs is on the horizon, but they are only as good as the quality of the end points. The voice of the sirens that call us to novelty may be fraught with rocky shores and rough seas, as our Greek Mythology warns.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Novel designs

- Effective treatments for multiple sclerosis (MS), coupled with earlier treatment from diagnosis, has resulted in the need for larger sample sizes for standard clinical trial designs.
- The magnitude and rising costs of clinical trials has led to an interest in novel study designs.
- The novel designs considered are those that have not been frequently used in MS trials to date.

Bayesian designs

- Bayesian designs allow for the use of instructive prior information towards estimates for power and sample size.
- Such designs can be used with noninformative priors, but may require larger samples sizes to achieve desired precision.
- Bayesian designs do not require multiple testing adjustments, but require more information during trial planning.

Adaptive designs

- Adaptive designs allow for readjustment of sample size based upon interim estimates of outcome and variation.
- Such designs allow for seamless transition from a Phase II to a Phase III trial, but the timing of adjustment is extremely critical.
- Few adaptive design clinical trials have been completed to date.

Executive summary (cont.)

Treatment strategy designs

- Treatment strategy designs allow for clinical trial designs that more closely mirror clinical practice decisions by allowing multiple treatments used consecutively or concurrently.
- Sample sizes are generally larger given the multiple treatment strategy arms and heterogeneous nature of subject treatment patterns.

Growth modulation index designs

- Growth modulation index designs are defined as the ratio of time to progression on a new treatment compared with time to progression on the subject's most recent prior treatment – each subject serves as their own control.
- Progression and failure criteria are not well-defined in MS, and the number of available treatments is currently small.

Random initiation designs & withdrawal designs

- Random initiation designs and withdrawal designs stagger groups of patients to be started or withdrawn from drug(s) at random times to determine if there is a time-dependent result.
- Such designs may be useful with combination therapies, but few studies have been conducted in MS.

Futility designs

- Futility designs allow early termination of a trial which has a low probability of achieving success.
- Such designs are cost-effective, but by design will miss effective treatments that have a delayed effect.

Future perspective

- The advent of additional treatments will allow for expansion of trial designs in MS.
- Regardless of the design, study results ultimately depend upon a quality end point that is clinically and statistically relevant.

Bibliography

Papers of special note have been highlighted as of interest (•) or of considerable interest (••) to readers.

1. Von Hoff DD: There are no bad anticancer agents, only bad clinical trial designs – twenty-first Richard and Hinda Rosenthal Foundation Award Lecture 1. *Clin. Cancer Res.* 4(5), 1079–1086 (1998).

• **Highlights the importance of the trial design reflecting the mechanism of action of the agent being tested.**

2. Cutter GC, Cofield SS: Outcome Measures in MS. In: *Multiple Sclerosis*: A Comprehensive Text. Raine CS, McFarland HF, Hohlfeld R (Eds). Saunders, Elsevier, Edinburgh, UK, 447–456 (2008).
3. Sormani MP, Rovaris M, Bagnato F *et al.*: Sample size estimations for MRI-monitored trials of MS comparing new vs standard treatments. *Neurology* 57, 1883–1885 (2001).
4. Sormani MP, Miller DH, Comi G *et al.*: Clinical trials of multiple sclerosis monitored with enhanced MRI: new sample size calculations based on large data sets. *J. Neurol. Neurosurg. Psychiatry* 70(4), 494–499 (2001).

•• **Utilization of simulation methods to estimate sample sizes needed for MRI trials, specific to multiple sclerosis.**

5. Cutter GC, Cofield SS, Sychowski J, Conwit R, Lublin FD, Wolinsky JS: Analyzing Relapse Assessments in CombiRx. Presented at: *23rd Congress of ECTRIMS*, Prague, Czech Republic. 11–14, October 2007.
6. Cutter GR, Baier ML, Rudick RA *et al.*: Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* 122(5), 871–882 (1999).
7. Ameer B: Novel trial design: a report from the 19th Frontiers Symposium of ACCP. *J. Clin. Pharmacol.* 48(7), 793–798 (2008).

• **A discussion of Bayesian and other approaches that may lead to better decision-making in the design and execution of clinical trials.**

8. Mick R, Crowley J, Carroll R: Phase II clinical trial design for noncytotoxic anticancer agents for which time to disease progression is the primary endpoint. *Control. Clin. Trials* 21(4), 343–359 (2000).
9. Suman VJ, Dueck A, Sargent DJ: Clinical trials of novel and targeted therapies: endpoints, trial design, and analysis. *Cancer Invest.* 26(5), 439–444 (2008).

• **A discussion of modifications to traditional Phase I, II, and III clinical trial designs. Advises using multiple end points in earlier stage trials to refine later stage trial designs.**

10. The NINDS NET-PD Investigators: A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. *Neurology* 66, 664–671 (2006).

•• **Assessment of futility of agents in a clinical trial, resulting in consideration of other factors prior to the selection of agents for clinical trials.**

11. Tilley BC, Palesch YY, Kieburtz K *et al.*: Optimizing the ongoing search for new treatments for Parkinson disease: using futility designs. *Neurology* 66, 628–633 (2006).
12. Schwid S, Cutter G: Futility studies: Spending a little to save a lot. *Neurology* 66, 626–627 (2006).