ADAPTIVE BAYESIAN DESIGNS FOR DOSE-RANGING DRUG TRIALS

Donald A. Berry[1], Peter Mueller[2], Andy P. Grieve[3], Michael Smith[3], Tom Parke[4], Richard Blazek[3], Neil Mitchard[3], Michael Krams[3]

[1]Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030. [2]Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251. [3]Pfizer Central Research, Clinical Development, Sandwich, CT13 9NJ, England. [4]Tessella Support Services plc, 3 Vineyard Chambers, Abingdon, OX14 3PX, England.

Table of Contents

1. INTRODUCTION

1.1 Stroke and neuroprotective agents

If you hold this text in your right hand and read it aloud, parts of the left side of your brain would be very active. This activity would consume considerable amounts of energy. Arterial blood provides a constant source of fresh energy to the brain in the form of glucose and oxygen. Now imagine that there is a sudden interruption in the arterial flow to these parts of your brain. Brain cells would start to die within minutes. You would lose control and power in your right arm and the text would drop. You would want to say something but would be unable to speak, and you may be unable to understand what other people are saying to you. These effects would be immediate.

Each year in the United States some 700,000 patients experience such symptoms. A doctor would diagnose "stroke" or "brain attack". In industrialized countries, stroke is the leading cause of disability, and it is the third leading cause of death, behind only heart disease and cancer. Stroke survivors can show either full or partial recovery. At about three months after the event most patients have reached their best recovery and at most small improvements occur after three months.

The overall economic burden of stroke is enormousÑapproximately $41 billion in the U.S. in 1997.

Until recently, there were few therapeutic options for stroke patients. This is now changing: both patients and medical teams have learned to deal with a brain attack as a medical emergency. The motto is ÒTime is brainÓ.There are two therapeutic approaches that have been shown to improve outcome: 1) Stroke units: Specialized semi-intensive care units, where blood pressure, blood sugar and other important factors relevant for outcome can be properly controlled and where patients receive physiotherapy. 2) Thrombolytic therapy: If a patient suffering from an ischemic stroke is admitted to a stroke unit within 3 hours of onset of symptoms, tPA (tissue plasminogen activator) can help to dissolve the blood clot that is occluding the cerebral artery. tPA can reperfuse the affected parts of the brain (that is, get the blood flowing through them again). However, tPA has risks: given to the wrong patient, it can lead to serious complications, such as secondary hemorrhage.

A third therapeutic concept involving neuroprotective agents is under development but is still experimental. The example considered throughout this report is the development of a particular neurological agent. The logic of neuroprotective therapy in acute stroke is based on the pathophysiology of brain attacks. The following is a brief review of that pathophysiology.

Brain cells cannot function properly without a constant supply of oxygen and glucose. Stopping blood flow halts this supply. An interruption for longer than a few minutes will result in the death of affected brain cells. Once cells start to die, the fine equilibrium of intra- and extracellular ions spirals out of control. Neurotransmitters excite surrounding brain cells. This makes them hyperactive in an environment where the energy supply is

already at dangerously low levels, which leads to further cell death. Complex inflammatory processes are triggered and complicate the situation further. The area of the brain that receives no or minimal blood flow is called the *infarct core*. Cells in the infarct core are destined to die. The *penumbra* is the area of the brain that surrounds the core and that receives some blood flow, but at a level insufficient to allow brain cells to function properly. Cells in the penumbra are able to survive. After a time they may recover some or all of their function.

Think of a stroke as a rapidly expanding sphere inside a brain. The interior of the sphere (the infarct core)ÑbutÑnotÑitsÑsurfaceÑ(penumbra)ÑisÑcondemnedÑtoÑdeath. Now imagine that it is possible to intervene and interrupt the expansion of this sphere. Fewer brain cells would die. The goal of neuroprotective therapy in stroke is to stop or slow the expansion of the infarct core. To be effective it may have to be administered very soon after the initiation of stroke.

Some neuroprotective agents have been tested in animal stroke models. These tests are controlled for important factors such as:

- Type of stroke (ischemic, cortical/subcortical).
- Stroke severity (length of time over which a cerebral artery is occluded—the longer the occlusion, the more severe the infarct).
- Time delay between onset of stroke and administration of therapy (the shorter the delay, the more efficient the therapy).
- Administration of study drug (dose and duration over which drug is given).

Evidence from animal experiments suggests that neuroprotective agents can reduce infarct volume. However, whether they do so in humans and whether such reduction translates into clinical benefit are not known.

Neuroprotective agents that are promising in animal models may be tested in clinical trials. Generally in drug development, the first clinical trials are in human volunteers and the goal is to assess the agentÕs safety. If it is proven safe then it is used in patients who have suffered stroke. The goal is to determine whether the agent is effective, and of course to continue assessing its safety.

Assessing neuroprotective drug efficacy is not easy. The size of a patientÕs infarct core cannot be observed directly. There are sophisticated imaging techniques (diffusion/perfusion weighted magnetic resonance imaging) that measure infarct volume and change in infarct volume over time. But the same image can result from quite different stroke kinds and severities. Neurological deficit resulting from a stroke depends on a) the stroke locale, b) the extent of the brain damage, and c) the strokeÕs etiology. A large cortical lesion in the primary motor cortex can produce a neurological deficit similar to a very small lesion along the corticospinal tract. And a rather large lesion in certain parts of the frontal lobes may not produce only minor deficits, or least deficits that

are difficult to measure. As regards etiology, approximately 80% of strokes are called "ischemic", involving the obstruction of a cerebral artery. And about 15% of strokes are "hemorrhagic": if an arterial vessel bursts, blood will be pushed with high pressure into the delicate brain tissue, destroying the fine neuronal structures. Sometimes a hemorrhagic stroke can be a secondary complication of an ischemic stroke. This can happen when the lack of blood perfusion mollifies the brain tissue. When the occluded artery becomes cleared, blood may shoot with high pressure into the area of the brain that had been damaged, producing a "reperfusion injury", and potentially resulting in secondary hemorrhage.

Sophisticated imaging methods are not widely available. Even if they were common and even if they were perfectly accurate, size of infarct may not be very predictive of the stroke patient's recovery.

In practice, most stroke trials take the principal endpoint to be the patient's condition 90 days after the stroke. The patient's condition is measured by one or more outcome rating scales. There is no general agreement regarding the appropriate way to assess recovery in stroke patients. Instruments are difficult to standardize across different types of strokes, different cultures, and different expectations. Some stroke scales for assessing recovery are dichotomous, with patients classified as "dependent" or "independent". Other scales have a finer rating of neurological deficit. We will use one of the latter, the Scandinavian Stroke Scale (Scandinavian Stroke Study Group 1985). The SSS has nine components, each with possible scores ranging from 0 to some maximum number of points. The maxima are shown in parentheses in the following list: consciousness (6), eye movement (4), upper extremity (6), hand (6), lower extremity (6), orientation (6), aphasia (10), facial (2), walking (12). Higher score means better performance. The maximum total SSS is 58.

Inclusion/exclusion criteria in stroke trials are very important, although exactly what they should be is not clear. Patients with mild stroke are easier to recruit into trials, in part because there are more of them. Such patients tend to improve even without therapy. On the other hand, perhaps they would improve even faster with neurological therapy.

To help us understand the response over time of untreated patients we used an extensive longitudinal database over time for in excess of 1000 acute stroke patients, the Copenhagen Stroke Database (J¿rgensen, et al. 1994). Perhaps the most impressive aspect of this database is that it demonstrates the difficulty of developing drugs for treating stroke: the variability in response is enormous, even considering patients with similar baseline characteristics, including baseline SSS. Information from this database led us to focus on moderate to severe stroke in the trial described in this report. In addition, we use this database to simulate trials, as described later in this report.

Among the most important aspects of drug development programs is selecting the correct dose. To many programs have focused too little attention to this aspect of drug development. There have been a number of failures in recent trials investigating neuroprotective agents in acute stroke and these may be due to an imperfect understanding of the dose-response relationship and a consequent incorrect selection of dose.

One of the reasons for the inadequate understanding about dose-response in development programs of neuroprotective agents is that proper learning requires trials with a very large number of patientsÑat least with traditional designs. Given the time and resource constraints, some programs may have extrapolated dose-response relationships from preclinical programs, sometimes assuming that from the perspective of efficacy, "more is better".

Our approach described in the following section considers only single-shot IV therapy to be given as soon as possible after the stroke. We do not consider duration of therapy as an independent factor. A more comprehensive model would allow for learning about this dimension of Òdose.Ó

We have suggested that acute stroke is a difficult indication for which to develop therapies. The design we propose in this report improves on the efficiency of drug development programs by applying Bayesian thinking and adaptive design. This methodology makes learning about novel therapies more efficient. It will help in the treatment of stroke patients by reducing the numbers of patients exposed to non-efficacious doses or drugs during clinical development of a new compound and by bringing efficacious therapies to the marketplace earlier. The approach is applicable to a wide range of drugs and indications, with stroke being but an example.

1.1. Critical Aspects of Standard Dose-Selection Trials

In a standard dose-selection trial, patients are assigned randomly to a set of pre-defined doses, typically between 3 and 5 in number. A standard trial is balanced in the sense that equal or nearly equal numbers of patients are assigned to the various doses.

Such a trial may be inefficient. If the test drug is effective then the dose-response curve has a positive slope for some interval of doses. The sloping part of the curve may be located at doses greater than or less than those considered in the trial. In either case some of the observations at the opposite end from where the slope occurs may be wasted. Moreover, if the sloping part of the curve occurs within the range of doses considered and is steep relative to the interval between doses then observations at both ends of the range may be largely wasted. The case in which the drug is ineffective is similar to the one in which a positive slope occurs at doses larger than those considered in the trial; namely, many of the observations at lower doses may be wasted.

To illustrate some of these points, consider the sample of dose-response curves shown in Figure 1.1, (a)-(f):

- The true dose-response curves are logistic.
- Four active doses (labeled 1, 2, 3, 4) plus placebo (dose 0) are assigned in a balanced fashion so the sample sizes are equal.
- The true values of the response curves at these doses are shown as solid circlesÑthese are the true mean responses for the various treatment arms in the trial.

- The vertical lines attached to the circles represent the set of likely values for the *observed* mean response at the corresponding dose. For example, this interval may be the one containing the middle 95% of the probability.

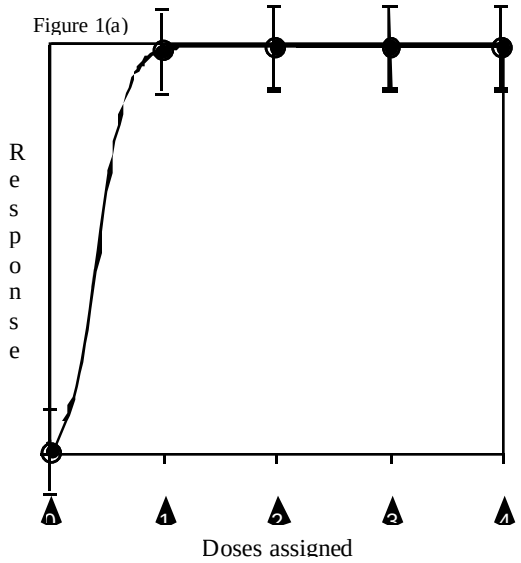


Figure 1(a)
Response
0
1
2
3
4
Doses assigned

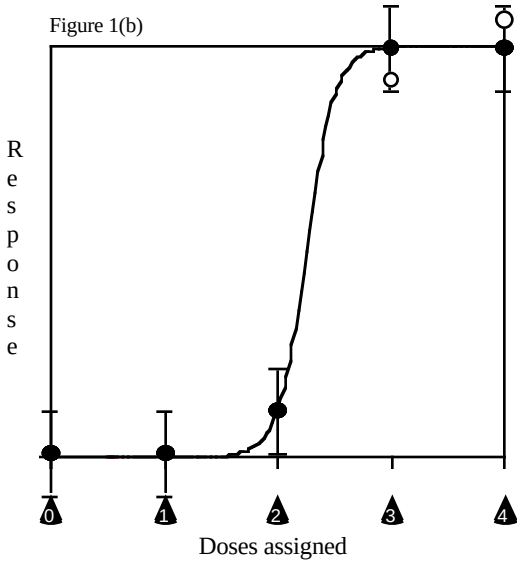


Figure 1(b)
Response
0
1
2
3
4
Doses assigned

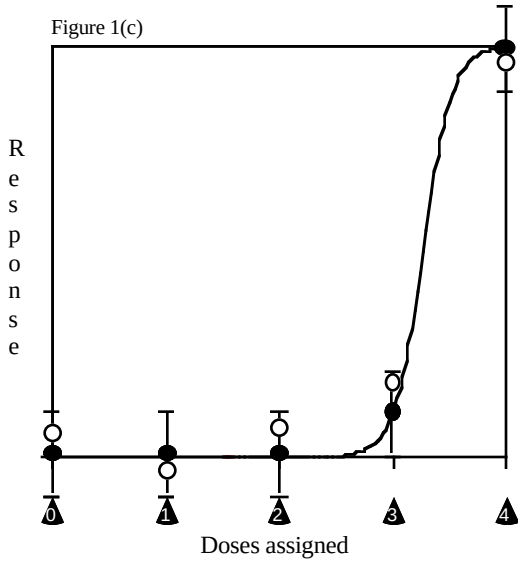


Figure 1(c)
Response
0
1
2
3
4
Doses assigned

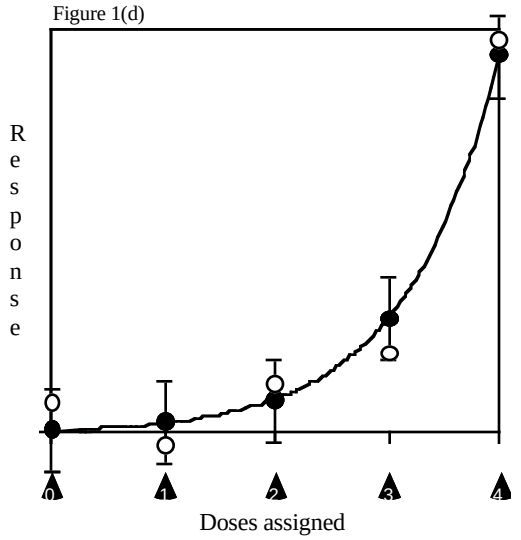


Figure 1(d)
Response
0
1
2
3
4
Doses assigned

Figure 1(e)

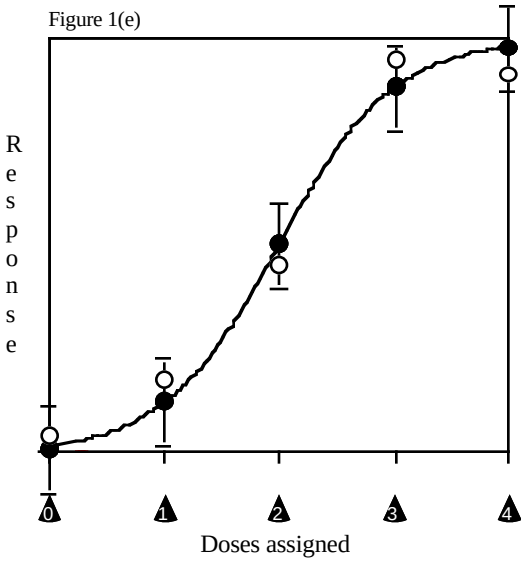
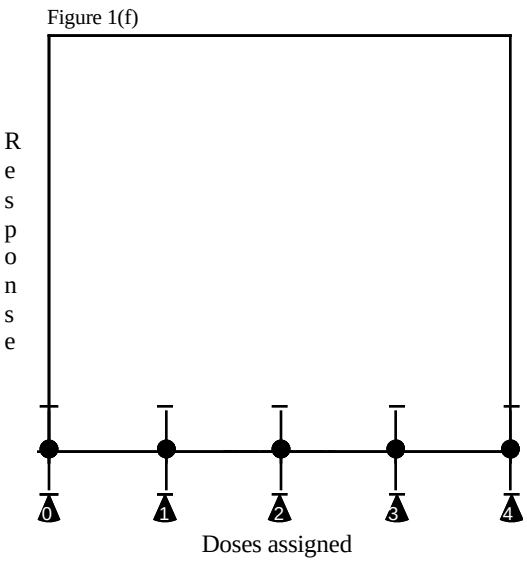


Figure 1(f)



In Figure 1.1(a) the sloping part of the dose-response curve is between doses 0 and 1, with all four active doses having essentially the same underlying mean response. In retrospect, many (but not all) of the patients assigned to doses 2, 3 and 4 would have been more informative about the slope of the dose-response curve (and also about the ED95, the smallest dose at which 95% of the maximal response is achieved) had they been assigned to doses between 0 and 1.

The accuracy with which a trial can determine the ED95 is limited by the separation between doses. In Figure 1.1(a), for example, the ED95 will likely be identified as being between doses 0 and 1. Such accuracy may be sufficient for some compounds, but for drugs that may be less safe at higher doses and for drugs that are costly to produce, it may be essential to identify the optimal dose with greater accuracy.

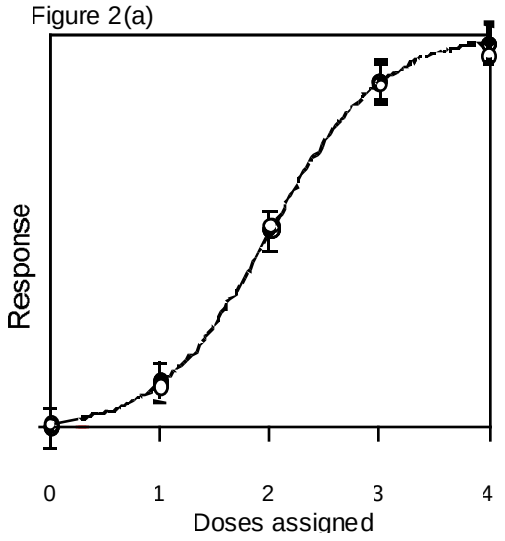
In Figure 1.1(b) the slope of the dose-response curve is the same as that in Figure 1.1(a) but the entire curve is shifted to the right. If the dose-response curve happens to be the one in Figure 1.1(b) then in retrospect it would have been more informative to have assigned patients to doses between 2 and 3 and also additional patients at doses 3 and 4. In addition, sampling variability present in the observed responses at doses 3 and 4 will leave open the possibility that the asymptote of the curve has not been reached. For example, the observed means at doses 3 and 4 may be as indicated by the open circles. But regardless of the observations at these two doses, the value of the asymptote and the ED95 will be poorly known, and in retrospect assigning patients to doses higher than 4 (safety permitting) would have been more informative.

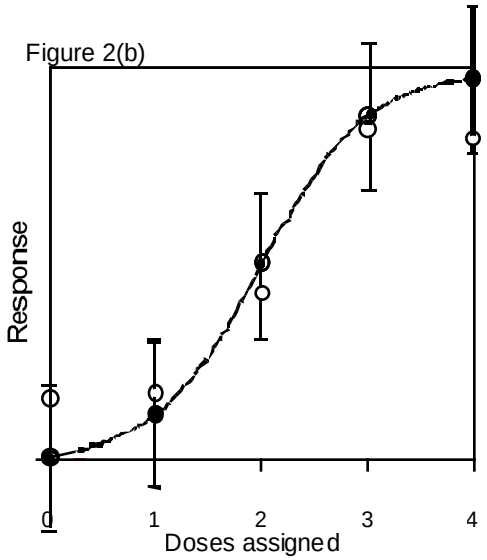
On this same point, the conclusions drawn from trials in which the curves are as shown in Figures 1(c) and 1(d) will be similar. For example, the five observed mean responses may be as shown by the open circles, and these are identical in the two figures. In 1(c) but not in 1(d) the asymptote will have been reached (unbeknownst to the investigators). Moreover, if the responses at dose 4 in comparison to placebo are not clinically important then in 1(c) the drug should be abandoned while in 1(d) doses higher than 4 should be explored, safety permitting. The observed results of the trial will not help in deciding which strategy is better.

In Figure 1.1(e) the doses assigned happen to be ideally placed in the sense that they encompass the gamut of responses and they are spread out along the sloping part of the curve. But even here, because of uncertainty concerning the true slope of the curve between doses 3 and 4, the ED95 is not very well identified.

In Figure 1.1(f) there is no response over the range of doses considered. Assuming that the dose-response curve is known to be monotonic, information from patients assigned to doses 1, 2 and 3 will be largely wasted and would have been more informative had they been placed at either end. Either the drug is ineffective or its effect occurs at doses higher than 4. Explorations beyond dose 4 are essential (if safe) to decide which possibility is the correct one.

Another drawback to standard types of designs is the use of fixed sample sizes. Even if the doses considered are judged in retrospect to be appropriate, the variability in response across patients assigned to the same dose may be less or greater than anticipated. In the former case the sample size chosen may be unnecessarily large and in the latter case it may be too small to accomplish the study's goals. Figure 1.2 addresses this issue. In both 2(a) and 2(b) the dose-response curve is the same as in Figure 1.1(e). In Figure 1.2(a) the measurement standard deviation is half that of Figure 1.1(e) and in Figure 1.2(b) it is twice that of Figure 1.1(e). The standard errors at each dose will then be half as large and twice as large, respectively, as anticipated.





Not surprisingly, a retrospective view of conventional dose-finding trials suggests that assigning to doses rather different from those actually assigned would have been more informative, or that a different sample size would have been more appropriate, or both.

At least two types of improvements to balanced designs with pre-selected doses are possible:

(1) Increase the number of doses considered. This can facilitate identifying critical aspects of the dose-response curve, but if the numbers of patients assigned at each dose is the same then the cost in terms of sample size may be prohibitively large.

(1) Avoid large sample numbers of observations at doses where the form of the curve can be reasonably well estimated from a smaller number by borrowing the information available from observations at nearby doses.

Both improvements are accomplished using the procedure we describe in this report. The result is that more accurate information is available about critical aspects of dose-response with, generally, smaller sample sizes. Moreover, when the total sample size is larger, this larger size is necessary to identify important aspects of the dose-response curve.

In this report we propose an adaptive scheme in which the doses assigned are allowed to depend on information from patients treated previously. The design is in two phases:

a) The focus of the first phase is dose finding—Is there a dose that is sufficiently effective to be used in a pivotal trial and thereafter in a clinical setting?

b) The second phase is confirmatory.

A second adaptive aspect of the design is that the duration of the dose-finding phase of the trial depends on the accumulating information about the dose-response curve:

- If the available information is sufficient to conclude that there is indeed a response to the drug and that the dose-response curve is well identified, assignment switches to a confirmatory mode in which a particular dose that has been identified as being effective—such as the ED95—is compared with placebo in a balanced, randomized fashion. The switch can take place seamlessly with no delay in patient accrual.

- On the other hand, if the available evidence is sufficient to suggest that there is no dose-response or that the maximal effect is too small to make the drug worthwhile clinically, then dose finding will cease and the procedure will recommend terminating the trial.

The design is adaptive in both assigning dose and deciding when to stop dose finding. And it uses Bayesian updating and Bayesian decision analysis. However, the regulatory approval process is largely frequentist. Therefore, the decision analysis involves predicting statistical significance (in the usual frequentist sense) of the confirmatory trial, and its sample size is determined accordingly.

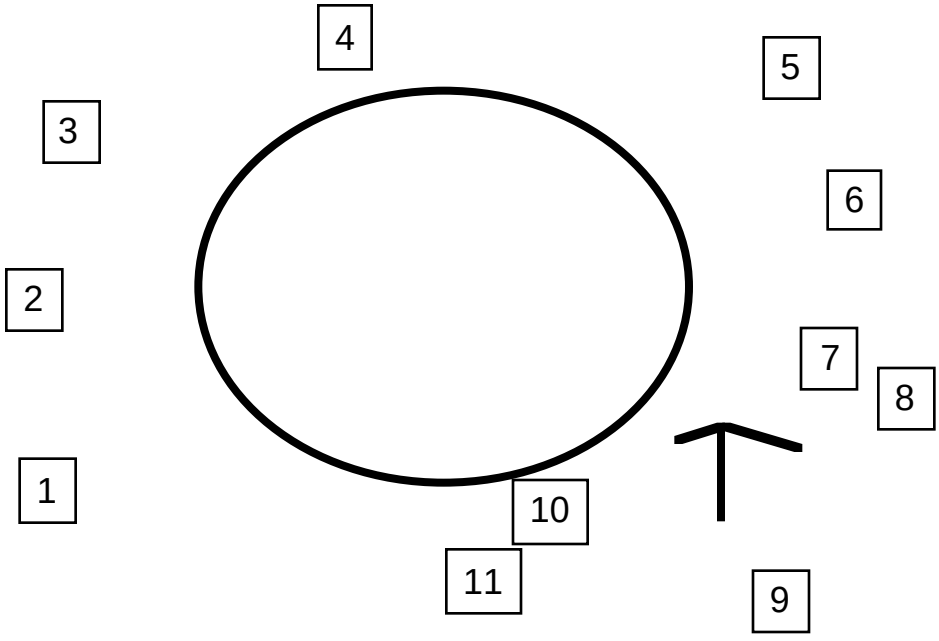
There is a regulatory tradition of requiring two trials, both with statistical significant results. The design we propose embodies two trials, one the dose-finding phase and the other the confirmatory phase. Statistical significance in the first phase can for the slope of the dose-response curve or for the comparison of the ED95 with placebo, assuming a particular probability model. The current version of our program does not require statistical significance in the first phase (adjusted for the sequential nature of the design) before converting to the second phase. However, this option is being incorporated into the program so as to meet this requirement if it is imposed by regulatory agencies.

2. DESIGN PROCESS

The proposed design for dose selection is in four parts, which by the nature of sequential, adaptive designs are not necessarily ordered in time. The process is schematically represented in Figure 2.4 [[there are no Figures 2.1-2.3]], the 11 components of which will be described below and related to the four parts of the dose selection process.

2.1. Allocation of patients to optimal dose

On a particular day during the trial, a number of patients, typically fewer than 5, will be admitted to the trial [1 in Figure 2.4]. Consider the first patient. Based on the available information our knowledge concerning the ED95 will be represented by a posterior distribution with some variance. Assigning a particular dose to the next patient and observing the response will effect a change in the posterior distribution of the ED95, and hence in its variance. We can predict how the chosen dose and the corresponding response of the patient will change the consequent posterior variance. The most informative dose [11]Ñsay z*Ñis the one with the smallest average variance of the ED95, where the average is with respect to the predictive probabilities of the various future observations, given the dose assigned. Before assigning the dose, however, some amount of randomization is employed [2]. First, if z* is not placebo then placebo is assigned with some minimal probabilityÑsuch as 15%. Second, consider the set of doses for which the expected response is within 10% that of z*. The dose assigned is selected randomly from within this set. Subsequent patients admitted on a particular dayÑup to 5Ñare considered as above, but conditioning on the previously assigned doses and averaging with respect to the predictive distributions of the responses at these doses.



This model allows an early update of knowledge about the dose-response relationship, which in turn allows for better choices concerning the doses to be allocated to future patients. In view of the critical importance of delivering neuroprotective therapy as quickly as possible, such updating requires rapid data processing. A rapid communications interface between the statistical system and the centers where the study is carried out has been developed. Details of the interface are given in⁄5.

2.3. Estimation of the dose-response model

We use a dose-response model based on a Normal Linear Dynamic Model (NDLM) described by West and Harrison (1997). This is essentially a piecewise linear model. It provides the necessary flexibility to encompass both monotonic and non-monotonic dose-response relationships. An additional advantage of NDLM is the existence of analytical results for the determination of the posterior distribution of the dose-response curve, and also for the ED95. Together with data from those patients who have completed the study, the imputed values described in the previous section are used to update the relevant posterior distributions. Details of the model and the updating procedure are given in⁄6.

2.4. Confirmatory decision

On each day of the dose-finding phase the algorithm implements a decision rule [7 in Figure 2.4] that addresses the question of whether the process should continue in dose-finding [10 in Figure 2.4], end because of futility [8 in Figure 2.4], or shift to a confirmatory phase [9 in Figure 2.4]. These decisions are made based on the predictive probabilities of failure to show a benefit and successfully showing a benefit. (The decision to stop the trial is not made in the absence of human intervention. A Data Safety Monitoring Board reviews any decision concerning the conduct of the trial.) The shift from the dose-selection finding to confirmatory may be seamless in that the phase of the trial may not be recognizable by physicians and others involved. If a shift is made in a seamless fashion then the follow-up information from patients treated most recently in the trial will not be complete.

The timing of the shift from dose-finding to confirmatory is important. Whether to shift is based on a Bayesian decision analysis. A component of this analysis is an assessment of the probability of showing an eventual benefit in a confirmatory trial at the current ED95. The decision to shift will depend on the available information about dose-response. The design of the confirmatory study will be based on the predictive power calculations, but whether to shift to the confirmatory phase of drug development is based on Bayesian calculations of the probability of showing a benefit over placebo.

Insert either cmu.pdf or cmu.ps here

These inserts contain the following sections [pardon the inconsistent numbering].

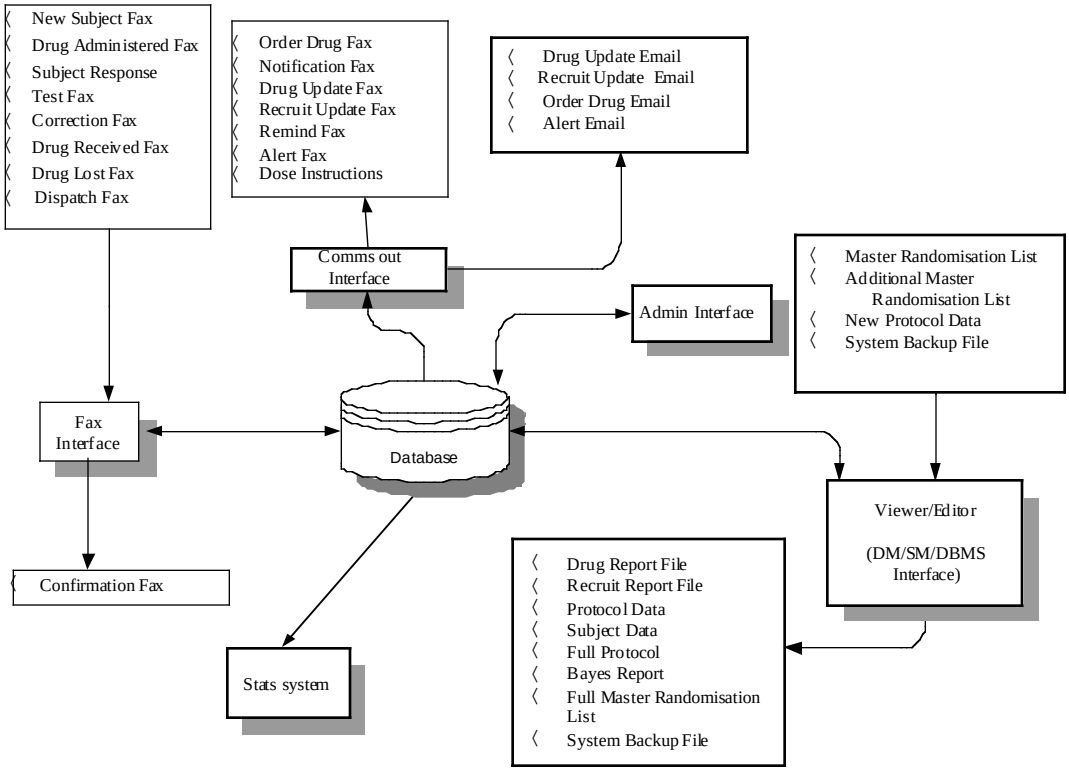
2. **The System**
3. **Dose Allocator**
4. **Optimal Stopping: The Terminator**
5. **The Dose-Response Curve: Probability Model**
6. **The Longitudinal Model**

7. INVESTIGATOR/SYSTEM INTERFACE

7.1 Introduction

The Response\01 Data System makes it possible to use a Bayesian Adaptive Design in a Clinical Trial. It receives the information from the investigators, orchestrates the running of the statistics programs, transmits to the investigators the details of what dose to administer the patient, and provides a means of monitoring the progress of the study.

System Overview



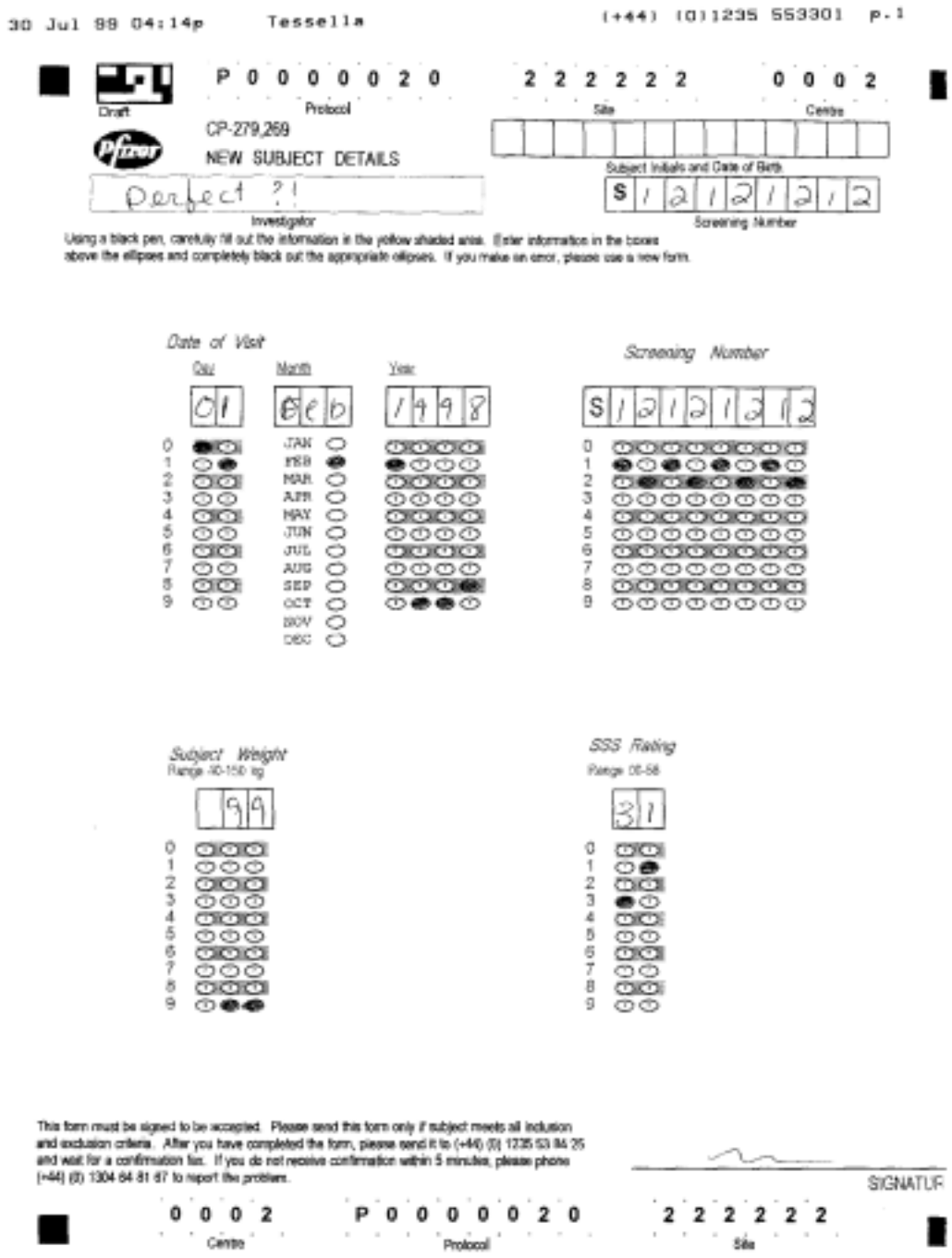
7.2 Using the System—The Investigator

Upon admitting a patient suitable for the study the investigator:

1. assesses their weight and their baseline score on the Scandinavian Stroke Scale,
2. enters this information and the patient initials and date of birth on a single page pre-printed form,
3. faxes the form to data system,
4. receives a confirmation of her fax, with the number for the patient within the study, and the details as read from the investigator's form by the system so they can be checked,

5. a minute or two later, receives a fax with instructions on how to make up the dose to give the patient.

An example fax:



The dose assigned to a patient is specified to the investigator, or the investigatorÕs pharmacy, by identifying two vials and instructing how much is to be taken from each and combined to make the dose. A unique number identifies each vial. Vials contain one of three concentrations of the drug or placebo.

This allows a wide range of doses to be used and larger than usual number of different uses within the range is available for the dose allocation algorithm. The system is also able to scale the dose according to the patientÕs weight.

As well as the fax forms with their minimum of information, the investigators complete more detailed study forms and pass them to the CRAs in the usual way.

To enable updating the current distribution of the relevant parameters, the investigators complete a form indicating the results of follow-up visits for all patients in the study. They also fax these to the system. Should a patient develop complications the investigator can request being unblinding as regards that patient, again by completing a form and faxing it to the system. The system replies with details of the dose assigned to the patient.

Faxed forms are being used rather than an internet and Web Browser-based interface because of the lack of availability of such facilities in some participating centers, and the fear that if the study sponsor provided the necessary facilities they would be abused or stolen. Faxed forms are slower and less reliable than an internet based solution, and implementing an interface would be easier if internet access could be taken for granted.

7.3 Using the SystemÑThe Study Administrators

The system will hold a large amount of relevant and up-to-date information of interest to the Study Administrators (Statisticians, Data Managers, Study Manager and the Data Safety Monitoring Board). This is made available to them through a conventional PC-based program that allows the data to be navigated and viewed.

<<screen shot of programÑnot yet available >>

As well as the data, all the fax images can be viewed. This allows the Data Manager to check the data on the form against that read in by the system and correct it where necessary. A number of events, e.g. invalid faxes, receipt of correction forms, and requests to unblind patients, are flagged to the Data Manger to check and make corrections if necessary.

The data are available more quickly and more accessibly than for conventional trials. An additional advantage of having such a system is that it can proactively send reminders to investigators. For example, it can remind them when follow-up checks are due on patients. And it can let them know when their pharmacy is running low so they can request further packs of vials.

The information on doses, effective doses and the status of the study as determined by the terminator algorithm are blinded to all but the Data Safety Monitoring Board.

7.4 Using the System—The Pfizer Pharmacy

The pharmacy at Pfizer is responsible for providing a Master Randomization List of all the vials that may be used in the study, their number, the concentration they contain and their expiration date. Whenever the pharmacy dispatches a shipment of vials to a center they send a fax to the system indicating the vials sent.

Thus, the system has the information it requires to specify doses to investigators in terms of the blinded vial numbers.

The system is informed that the vials are available at a center by the investigator faxing in a form acknowledging receipt of a shipment. The investigators also have a form they can fax should a dose administered be other than that directed and a form indicating that vials have been lost or broken.

The system sends a fax or e-mail to the Pfizer pharmacy when a center's supply of any particular concentration of drug is getting low. It also provides regular reports to the Pfizer pharmacy indicating what vials are at which center and which vials are in transit.

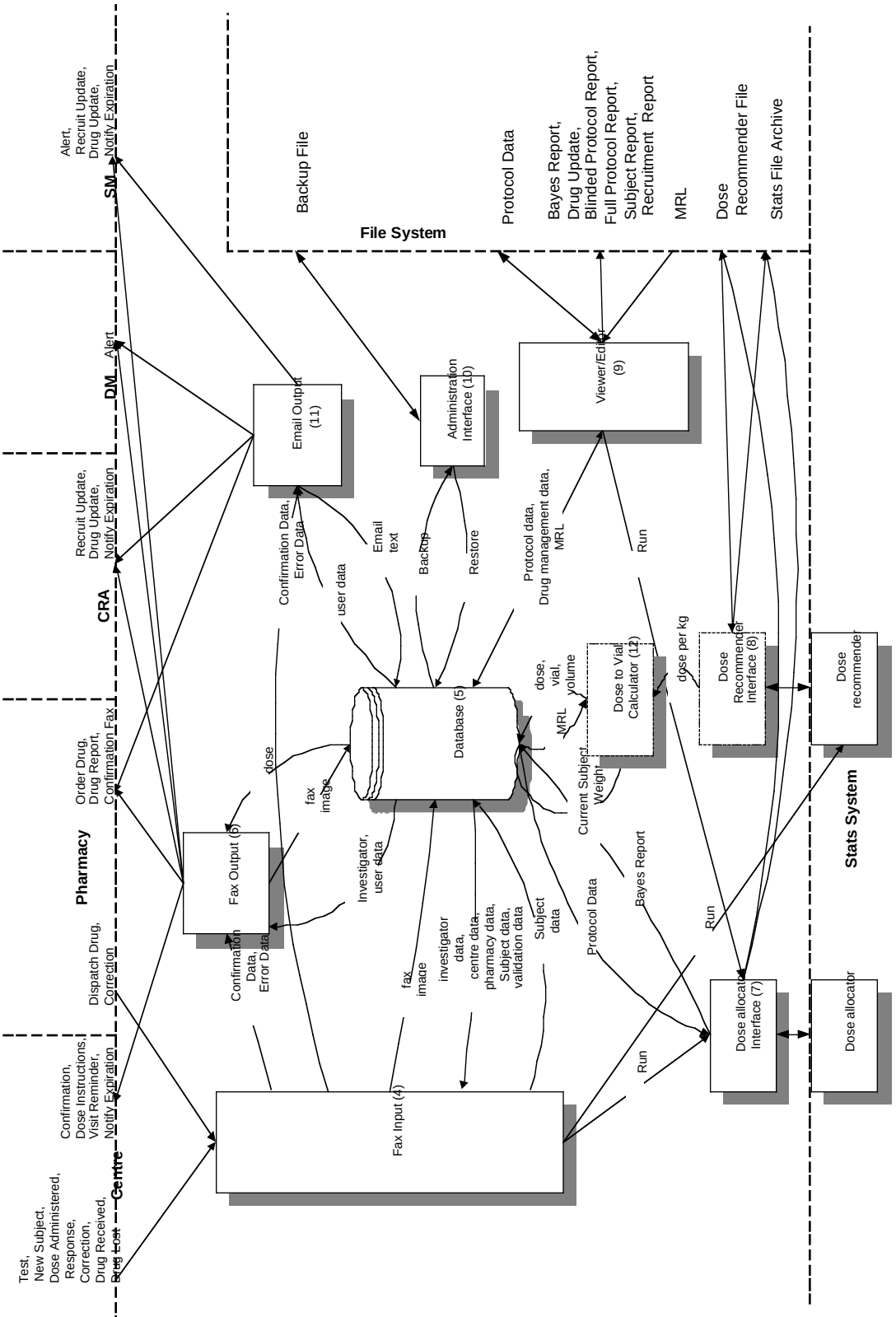
7.5 How the System is Constructed

The system comprises two servers and a number of connected client PCs. There is one Fax server and one Statistics server. The Fax server runs the Data System and the Statistics server runs the Statistics Programs.

The Data System comprises:

- a database built using Oracle 8,
- a fax reception package "Zetafax",
- the form recognition package "Teleforms" with the forms designed specifically for this system and the associated Basic scripts to validate the data and enter into the database,
- the Study Administrators client program, written in Visual Basic,
- the Fax-Statistics interface programs written in C and UNIX shell script.

The System dataflow model



7.6 The Development Process

The system has been developed in two phases.

1. The first phase concentrated on the core areas of:
    a) receiving patient data on faxes,
    b) storing the data in a database,
    c) passing the information to the statistics programs, running the programs and storing the results,
    d) providing facilities to browse and edit the data.
2. The second phase extends the first phase system by:
    a) sending dose information to the investigator,
    b) managing the vial data,
    c) extending the Study Administration facilities,
    d) providing proactive reports, warnings and reminders.

For each phase:

- Detailed Use Cases and User Requirements were drawn up that detailed what the system had to achieve.
- These were refined into System Requirements that detailed the functionality and properties of a System.
- A Design was then drawn up for the system, including a definition of the interface between the Data System and the Statistics
- The components identified in the Design written.
- Individual components of the system were tested.
- The whole system was tested.
- The system was then handed over for acceptance testing then user testing.

In the first phase it soon became clear that the Data System would be written before the fax form designs were finalized or the statistics programs were finished. So we decided to develop two versions during the phase, the first using early versions of the fax forms and statistics programs and the second only after the fax form and statistics programs for the first phase had been finalized. The first version was released after the component testing was complete but holding back some development effort for later rework and the system testing was held over until the second version was complete.

This had the advantage of allowing a thorough evaluation of the Study Administration client program in the first version and for the comments to be taken into account in the rework for the second version and in not duplicating any testing effort.

The first phase took about 18 person months of effort to develop. At the time of this writing the second phase is not yet complete.

The Validation Process

The system was validated by:

- Project reviews to ensure consistency of the requirements documents and the design.
- Code reviews.
- Component testing, a component test plan was drawn up and "light weight" component test scripts drawn up—stating what was to be tested but not how it was to be tested. The tests were then performed by the developers themselves.
- System testing, a system test plan was drawn up, enumerating the test scripts to be written and cross referring the test scripts to the system requirements. The system test scripts contained more detailed instructions on what steps to take to perform he tests and on the expected outcomes. Independent testers ran the system tests. The system testing also included some unscripted testing of the user interfaces by an independent tester, this proved particularly effective at finding faults.

8. SIMULATION RESULTS

8.1 Dose range, data properties and underlying curves

The number of doses used is arbitrary, with a greater number being better from the perspective of learning about the dose-response curve. But logistical considerations limit the number in practice. We used sixteen possible doses in all simulations: 0, 0.1, 0.2, up to 1.5. These doses are realistic for the actual trial but the set doses to be used is not yet final. Equal spacing is natural for the NDLM but unequally spaced doses can be accommodated.

We investigate the properties of our design by assuming a variety of dose-response curves. These were chosen to reflect different scenarios that might be expected in a dose-response study (Figure 8.1).

- Curve 1: Null or flat dose-response, corresponding to no drug effect.

The shape of the next three curves is logistic with the lower 4 doses having no improvement over placebo, and the highest 5 doses forming a plateau at the maximal drug effect.

- Curve 2: Minimal clinically relevant benefit over placebo (2 points on SSS) at the highest dose.

- Curve 3: Modest benefit over placebo (4 points) at the highest dose.

- Curve 4: Very large benefit over placebo (8 points) at the highest dose.

The next two curves reach their maxima earlier (Curve 5) and later (Curve 6) than those above.

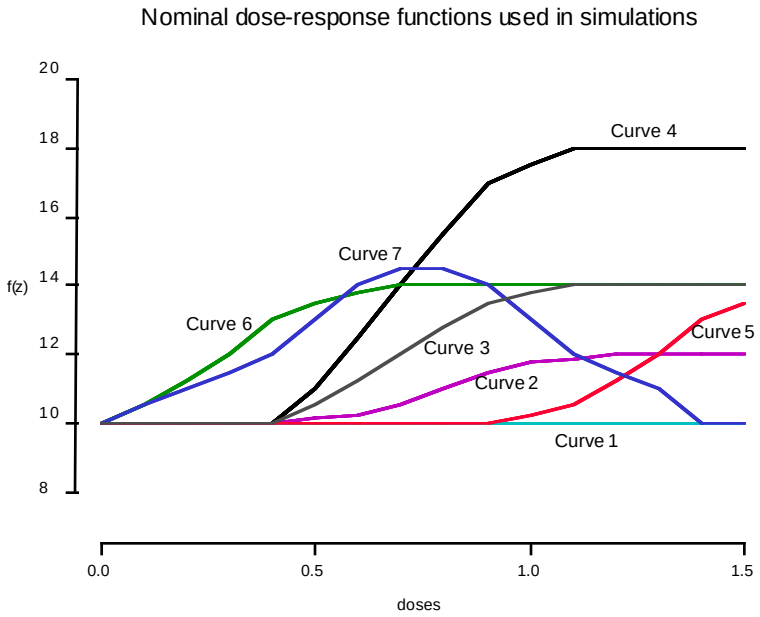
- Curve 5: Modest benefit over placebo (3.3 points) but none until the last 6 doses.

- Curve 6: Modest benefit over placebo (4 points) where the increasing is over the range of the first 6 doses. Similar to Curve 3 but with increase occurring 0.5 sooner in terms of dose.

The last curve we consider is the only one that is not monotonic.

- Curve 7: Bell-shape with increasing effect up to the middle doses and then the effect drops back to placebo level.

Figure 8.1ÑNominal dose-response curves used in simulations



In generating weekly patient data we use the Copenhagen Stroke Database (J¿rgensen et al. 1994). The SSS data for patients assigned to placebo are as in the CSD. A patient assigned to a positive dose responds as in the CSD, but with the improvement depending on the response at that dose for the curve assumed. For example, if Curve 3 or Curve 6 is assumed and the patient is assigned to dose 1.5 then the patientÕs change from baseline SSS is 4 points greater than actual.

Process and goals

We simulate a number of trials from a given curve. The statistical system incorporates software to enable us to generate patient data from the underlying dose-response curve, simulate the adaptive allocation of patients to doses, estimate the dose-response curve given the current patient data and make a decision regarding stopping the dose-finding phase of the trial.

We examine general properties of the algorithm, as follows.

- Doses allocated, depending on the underlying dose-response curve.
- Accuracy of estimate of dose-response curve.
- Sample size in comparison with standard designs.

We also wish to learn about the effect of different settings of the initialization files on the time taken to run the allocator and terminator functions. To use the system in an actual trial it must allocate a dose to the next patient in a matter of a few minutes.

There are a great deal of variable parameters in this approach and it is the aim of simulation to find out the effect of these parameters and identify optimal settings which allow best inferences to be made about the majority of the underlying curves.

Initialization file settings

An initialization file controls the critical aspects of running the algorithm. Appendices 1a and 1b contain details of the control items and their default values. For example, the default values for the prior estimates and longitudinal model analysis are from the Copenhagen Stroke Database. Default settings for Markov chain Monte Carlo (MCMC) iterations have been chosen to provide reasonable estimates of the parameters of interest but also to keep the calculation time within reasonable limits. Many options that can be set in the initialization file allow fallback positions should the software not function correctly. Default allocations include to the posterior quantiles and randomly to prespecified doses.

Other settings are invoked in generating data for simulations. In running a Ò trialÓ a patient database is created. If the settings indicate that the database should include deaths then this is incorporated. The patient data file is then read in sequentially (one week Õ s worth of responses at a time) and the doses are adaptively allocated to the patients in trial. The adaptive nature of the design means the simulation must be run with patients being allocated to doses and patient responses being generated sequentially.

The simulation process allows for setting the dose range to be used, the inclusion and exclusion criteria for baseline scores, the patient horizon (maximum number of patients), the maximum number of observations per patient and properties of the distribution and scoring method for deaths in the generated patient database. The initialization file also controls the specification of the prior knowledge about the form of the dose-response curve and controls some aspects of the eventual design of the study Ñ for example, the proportion of patients allocated to placebo.

In all simulations reported here we have set a lower limit of the sample size to be 250. This lower bound allows for getting at least a modest amount of data the safety of the drug. In addition, we set the upper limit on the sample size to be 1000. Some upper limit is helpful in carrying out the necessary calculations. Also, an upper limit is appealing to drug company management to facilitate budgeting and planning. The limit of 1000 is moderately large, as verified using simulation. When we set the limit to 5000, only occasionally did the terminator go beyond 1000 patients, irrespective of the dose-response curve assumed.

Simulation process

In most of the simulations reported here, 100 trials were generated for each curve. When initialization settings were changed, these were made one item at a time in order to investigate the effects and to keep the number of sources of change to a minimum at any one step. For example for any one nominal curve we simulated 100 trials with no deaths and with one setting of minimum and maximum sample sizes, one setting of placebo

proportion and one setting of the different NDLM parameters. Then we introduce deaths with a particular method of scoring death, holding all other factors constant. In this way we have attempted to build up a picture of the influence of different factors on our estimation of the underlying curve and on the inferences derived when we stop the trial.

To evaluate the overall and long-term performance of the algorithms and for comparison with standard designs we performed 1000 additional trial simulations on selected curves and for selected settings of the initialization file. These larger simulations are used to calculate Type I error probabilities (significant dose-response when the underlying dose-response curve is flat) and power of the procedure (proportion of trials with statistical significance for a nominal dose-response curve with the hypothesized clinically relevant difference).

The statistical package S-Plus was used to display the data from each simulation runÑwe examined the overall picture across the 100 trials and also observed the results for the individual trials.

Although we can evaluate the Type I error and power of our procedure there are very few metrics which will allow us to explicitly define its performance. As a result most of the evaluation from simulations has involved graphical presentation of the simulation results and comparing the dose allocation, estimation of the curve and stopping decision against what we would expect given knowledge of the underlying curve. We also compare the results of simulations against each other in order to identify any Òoddities.Ó

The conventional dose-finding study

We compare our results against conventional dose-finding study, one with three doses plus placebo. For a clinically relevant difference we wish to detect, for an estimate of variability, and for appropriate power and significance levels, we calculate sample sizes for each group. There are many assumptions underlying this calculation. The variability in the actual data may be greater than that assumed (taken from the Copenhagen Stroke Database). The effect of treatment may not be as great as assumed, or it may be greater. The location of the best effect may be outside the dose range assumed. And characteristics of the patient population may be different from those in other studies.

Any conventional study is saddled with such assumptions. If any are wrong then it may not be possible to draw a firm conclusion or the sample size may in retrospect be too large. Conventional studies may be able to recalculate variability and adjust sample sizes accordingly, or there may be other ways to adjust for some of the effects above while the study is ongoing, but in principle many factors are fixed at the outset of the study.

Table 1 shows sample sizes for a conventional study powered at 90% to detect a difference of 2 points or 4 points improvement over placebo. This assumes a 5% significance and testing using BartlettÕs correction for three active doses and placebo. Estimates of variance in change from baseline have been drawn from our analysis of the Copenhagen Stroke Database and reflect variability seen in an untreated patient population.

Table 1. Sample size estimates for a conventional dose-response study design.

Improvement over placebo to detect	Assumed variance; change from baseline	Total sample size (in 4 groups)
2 pts.	150	3216
4 pts.	150	804

Results of simulations

Consider Curve 3, where the maximal improvement over placebo is 4 points. The ED95 for this curve is approximately 1.0. Placebo is allocated with a fixed minimum proportion, in this case 0.15. One patient accrues per day. Figure 8.2 shows the frequency distribution of doses assigned in a single simulation (left panel) and the actual doses allocated by time over the course of the trial (right panel). The trial lasted 36 weeks and 250 patients were accrued. Therefore, at the time of stopping, complete information concerning the 13 weeks of the study was available on about 2/3 of the 250 patients.

Early in the trial the algorithm seeks to identify the minimum and maximum effects for these are critical references against which the ED95 must be evaluated. In the latter phases of the study the algorithm homes in on the ED95 value, while also getting information about the minimum and maximum. The histogram shows that a small fraction of doses are allocated in the dose range from 0.1 to 0.7. The nominal dose-response curve in this range is nearly flat and these doses are away from the area of principal interest, the region containing the ED95.

Figure 8.2. Results of a single simulation assuming Curve 3. Distribution of doses (left panel) and doses allocated over time (right panel).

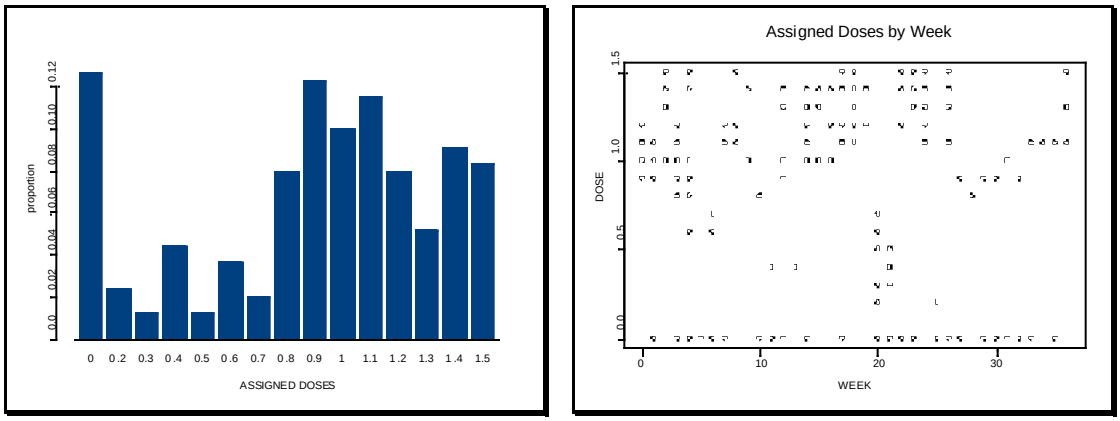


Figure 8.3 shows the distribution of doses across 100 simulations. Superimposed on this is a summary of the fitted dose-response curves across these simulations. The diamonds

represent the true (assumed) dose-response curve. The middle line shows the median of the fitted values with the quartiles of the distribution of the fitted values at each dose. This figure shows that the low doses have a low probability of being allocated. It also shows that pre-specified minimal proportion of placebo doses of 15% is achieved. The fitted curves accurately estimate the underlying dose-response curve.

Figure 8 3. Allocation of doses across 100 simulations, assuming Curve 3 for the dose-response relationship.

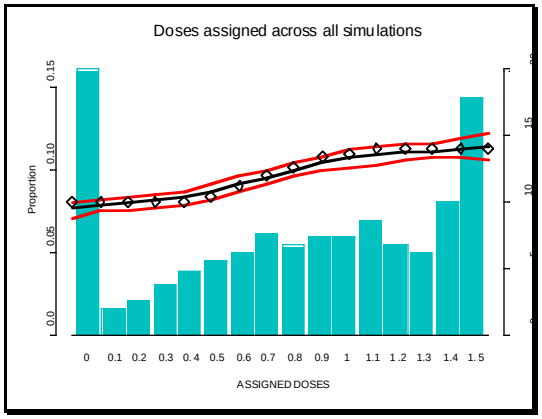


Figure 8 4 (left panel) shows the distribution of observed effect sizes (m) at the ED95 measured as the number of points improvement over placebo on the Scandinavian Stroke Scale against the standard error (s) of this estimate. The triangles represent those simulations in which the decision was made to go into a confirmatory phase and the circles correspond to simulations in which the decision is made to not go confirmatory. The sizes of the symbols are proportional to the trial sample sizes. Recall that the terminator function is called only after 250 patients have accrued and that the maximum study size is 1000. The figure shows that when the trial is stopped early it is because the effect is probably large. Naturally, for smaller sample sizes there is a bias in the observed response at the ED95.

For Curve 3, with its maximum of 4 points improvement over placebo the algorithm makes very few decisions to terminate the study because of lack of an observed benefit. In Figure 8.4) we see a "funnel" shape. This may be because in order to identify a small effect we will need to continue the trial to a larger number of patients and so we will end up with better precision in this estimate (smaller "s").

Figure 8.4 (right panel) shows distributions of the sample size taken to reach a decision using the stopping rule, the observed effect size, the standard error of the effect size and the posterior estimate of sigma. The bias in observed effect size at the ED95 can be seen from the distribution here, as can the fact that in the majority of cases the decision to stop has been taken comparatively early. The fact that some simulation runs have continued on to 1000 patients may be influencing the standard error of the estimated effect size

causing the bimodality we see in this distribution--the longer we continue the more certain we can be about the estimate of the effect size.

Figure 8.4. Distribution of observed effect sizes at the ED95 with corresponding precision. Dose-response is that of Curve 3. In the left panel symbol size is proportional to trial sample size at stopping. The right panel shows frequency distributions of sample size, posterior mean benefit over placebo at the ED95, its standard error, and the error standard deviation σ.

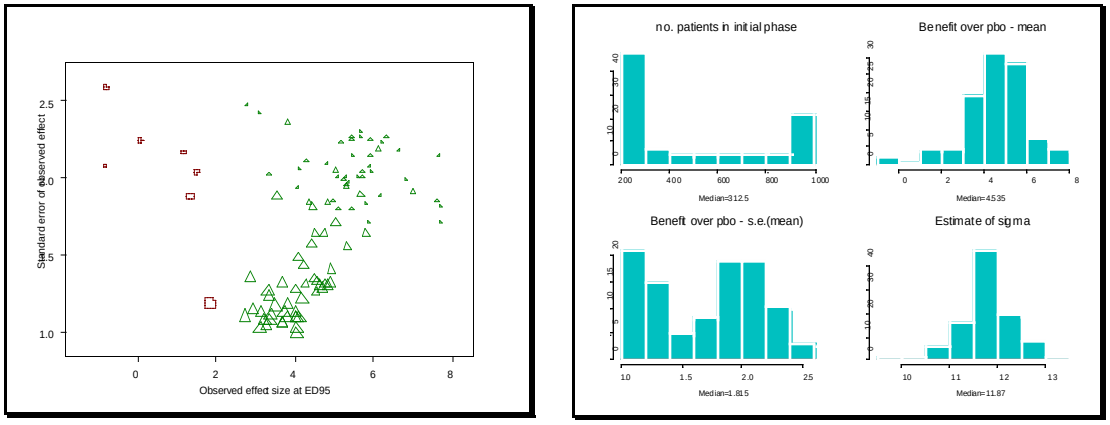


Figure 8.5. Distribution of observed effect sizes at ED95 with corresponding precision. The underlying dose-response is null, that of Curve 1. The symbol size in the left panel is proportional to sample size in the dose-finding phase.

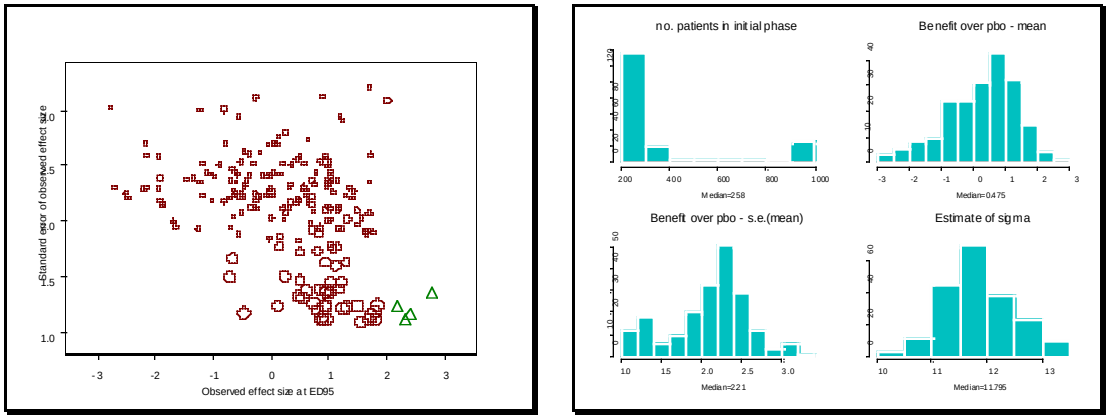


Figure 8.5 shows similar results for 200 simulations assuming the null or flat dose-response curve (Curve 1). The proportion of simulations in which the decision is to go into a confirmatory study is made is very low (2.2%). In the terminator function if the observed effect is less than 2 points improvement over placebo then we conclude that the

drug has insufficient clinical effect to warrant further development. The right panel of Figure 8.5 (upper left histogram) shows that the decision to stop was made after relatively few patients (median of 258 where the minimum is 250). As indicated in the introduction, an important benefit of our adaptive scheme is that when there is no drug effect the trial can be curtailed relatively early.

Figure 8.6 shows the distribution of doses for the remaining dose-response curves considered. The posterior estimates of the dose-response curves are superimposed on the histograms of doses. Evidently, our adaptive allocation scheme is capable of dealing with a variety of underlying dose-response relationships. In most cases it is effective in learning about the form of the underlying curve.

To find the statistical power of our adaptive scheme we calculate a minimal slope in the observed dose-response. ÒMinimal slopeÓ is the slope of a line spanning the whole dose range from placebo to the size of effect at the dose that will be selected to take into the confirmatory trial. This is minimal in the sense that if the maximal effect occurs anywhere other than at the maximum dose then the slope of the line will be greater. The slope is calculated only for those simulations that call for going forward into a confirmatory study. We can test this minimal slope for significance.

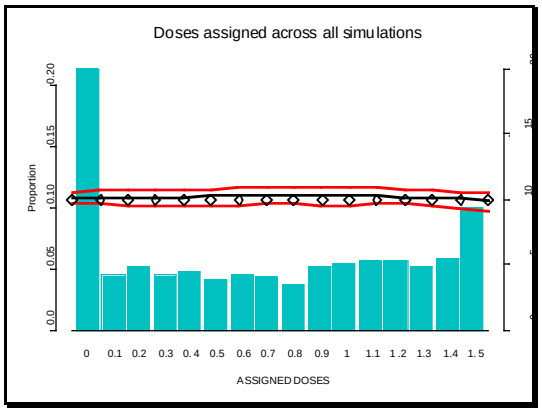
Table 2 gives a summary of the results from 100 simulations assuming Curve 3. In only 7 of 100 trials was the decision against going to confirmatory phase. As regards statistical significance of the minimal slope, only 3 out of the 93 simulated trials that go forward into the confirmatory phase did not show a significant slope.

The dose chosen from the adaptive scheme will be taken forward into a confirmatory trial. Since we know the true response at that dose we can easily calculate the power of the confirmatory study to show a significant difference from placebo. We can select a sample size for this confirmatory trial in a number of ways: using the effect size and estimate of variability from the dose-finding study, using a minimal clinically relevant difference along with the estimate of variability, and using predictive power to account for uncertainty in the estimate of effect from the dose-finding study. Estimates of the average power of the test for the confirmatory study using these three methods of sample sizing is also shown in Table 2.

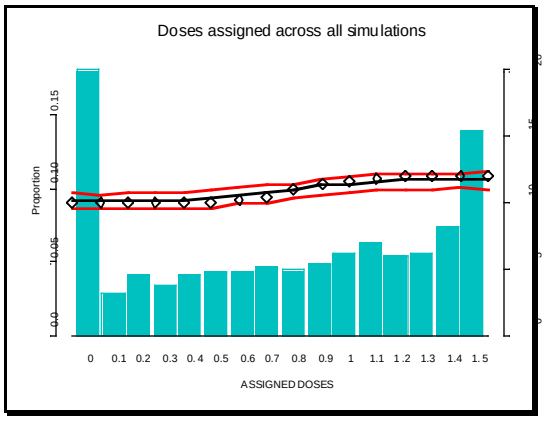
The power to show a difference from placebo in the confirmatory trial shows that if we use the minimal clinically relevant difference of 2 points improvement over placebo then we will overpower the study since the true effect is approximately 4 points improvement. Using the posterior estimate of the effect size at the ED95 leads to less than the nominal 90% power since we are not allowing for uncertainty in our estimate of effect size in the calculation of sample size. This is incorporated using the standard error around the estimate of effect size (as shown previously) and using predictive power methods. We can see that using this approach our average power is much closer to the nominal 90% power. Of course in practice all of these sample size estimates may be overridden by the need to have a substantial amount of safety data for regulatory submission.

Figure 8.6. Distribution of doses and estimated dose-response functions for dose-response Curves 1, 2, 5, 6 and 7 from Figure 8.1.

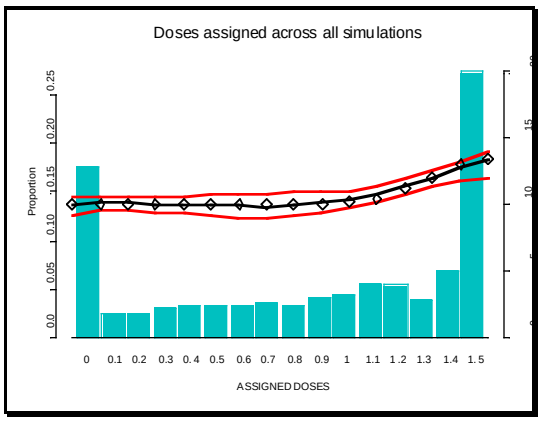
Curve˚1



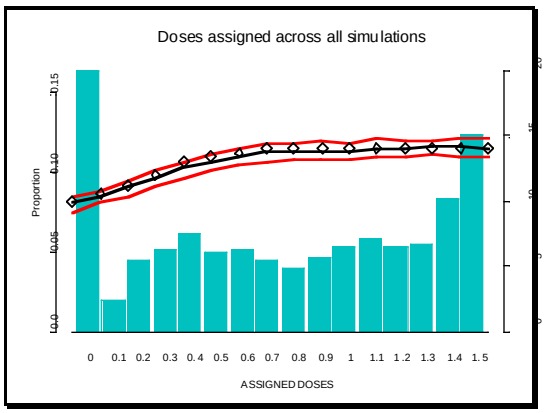
Curve˚2



Curve˚5



Curve˚6



Curve˚7

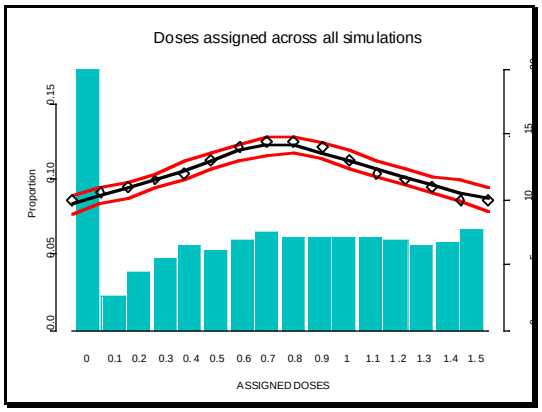


Table 2. Power of the adaptive approach assuming Curve 3 in which the maximal effect is 4 points improvement over placebo.

		Terminator decision	
		Stop (no effect)	**Continue to confirmatory trial**
Significance	**Not significant**	7	3
of slope *	**Significant**	0	90
	Total	7	93
Median sample size at stopping		255	343
Average power of test in confirmatory trial	**Posterior estimates of effect size, variability**		0.76 (108)
(median sample size)+	**Minimal clinically relevant difference (=2 pts)**		>0.99 (603)
	Predictive power		0.92 (174)

* Test of conservative estimate of slope = 0. ___

\+ Average power of confirmatory trial with z* (ED95 dose) and placebo to detect effect from known underlying dose-response curve at the chosen dose ***given*** that the dose-finding phase recommends going to confirmatory study. Nominal power=90%.

Table 3 shows the same summary for the null curve (Curve 1) in which there is no drug effect. This analysis allows provides the Type I error of this approach: any significant difference is a false-positive. In only 2% of the 200 simulated trials do we make the (incorrect) decision to go forward into a confirmatory study.

Table 3. Type I error of the adaptive approach compared to the conventional design. Curve 1--no effect.

		Terminator decision	
		Stop (no effect)	**Continue to confirmatory trial**
Significance of slope *	**Not significant**	196 (98%)	0
	Significant	0	4 (2%)
	Total	196 (98%)	4 (2%)
Median sample size at stopping		257	1000

* Test of conservative estimate of slope = 0. ____

Table 4 shows a summary similar to that shown above for Curve 3 but for different underlying variance. A major benefit of the adaptive design is that it adapts to the variability in the data. As a result, the confirmatory study will be powered appropriately based on the best estimate of variability from the data observed in the dose-finding phase. If variability in the data is substantially less than we would expect a priori then our inferences can also be made much sooner and with fewer patients.

However, Table 4 shows that when the variance is smaller, statistical power is not greatly increased. Rather, upon learning that the variance is small the algorithm stops the trial more quickly. The median sample size required to reach the decision to go forward into the confirmatory trial is smaller, and in fact is close to the minimum possible sample size of 250.

These power calculations show that again the adaptive design achieves close to the nominal 90% power for the confirmatory trial for either assumed variance. The predictive power approach takes into consideration the uncertainty in the estimate of the treatment difference and so achieves better power overall, albeit with larger average sample size.

Table 4. Power of the adaptive approach vs. conventional design. Curve 3Ñmaximal effect = 4 points improvement over placebo, different observed variances.

		Terminator decision	
		Stop (no effect)	Continue to confirmatory trial
Significance of slope *	Not significant	8	92
	Significant	0	0
	Total	8	92
Median n		1000	257
Power of test in confirmatory trial + (median n)+	Posterior estimates; effect size, variance		0.84 (42)
	Min. clin. relevant difference (=2 pts)		>0.99 (213)
	Predictive power		0.93 (60)
Significance of slope *	Not significant	14	3
	Significant	0	83
	Total	14	86
Median n		277	413
Power of test in confirmatory trial + (median n)+	Posterior estimates; effect size, variance		0.67 (150)
	Min. clin. relevant difference (=2 pts)		>0.99 (847)
	Predictive power		0.84 (238)

* Test of conservative estimate of slope = 0. ____

+ Average power of confirmatory trial with z* (ED95 dose) and placebo to detect effect from known underlying dose-response curve at the chosen dose ***given*** that the dose-finding phase recommends going to confirmatory study. Nominal power=90%.

From these simulations we conclude that our sample sizing and final inference across both trials is appropriate. The algorithm does what it is supposed to do: adapt to the underlying dose-response curve and to the variability in accumulating data. The dose-finding trial allows for stopping the drug development early either because it has no effect, or because the evidence is sufficiently clear to conclude that the treatment is effective. The simulations also show that when using the results of the dose-ranging phase to sample size a confirmatory trial, the latter will very likely confirm that the dose effect seen in the earlier trial is significant.

The simulations summarized above are (loosely) based on information from the Copenhagen Stroke Database but ignoring deaths. There are a number of issues we face in incorporating deaths into the generated data set, and we must deal with these as well when handling deaths in our analysis of the primary endpoint, change in SSS. Death must be included in some way.

To include deaths in the generated data we specify the distribution of deaths for different baseline scores—patients with mild initial stroke are less likely to die. We must also address the relationship between dose and death. The former we may evaluate from our analysis of the Copenhagen Stroke Database assuming that dose levels do not interact with baseline severity. The latter we must hypothesize at present because we have no information about whether and how the drug may affect the incidence of death. It may be that the effect of dose on death is not monotone, especially when averaging over severity of stroke. Perhaps low doses have no effect on mortality, medium doses offer protection against death and higher doses are linked to a mechanism that hastens death. This would give us a U-shaped incidence of deaths across the dose range.

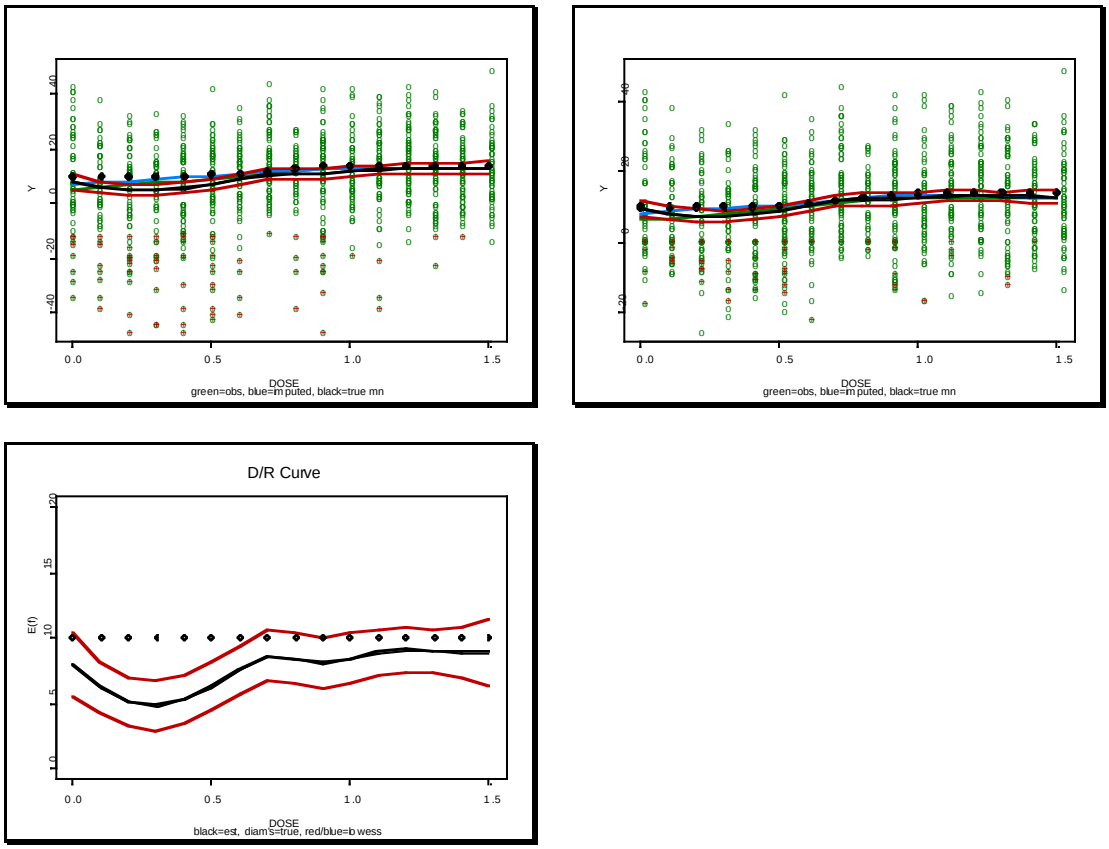
We want to incorporate death into the SSS endpoint, down-weighting any benefit of the treatment at doses when there is an increase in mortality rate. This will allow for addressing the role of mortality in the dose-response relationship. An alternative is to evaluate the dose-response among survivors only and analyze deaths separately. But the two analyses—SSS improvement by dose and death by dose—must be combined in some fashion and at some point.

It is far from clear how to incorporate deaths into the principal endpoint and as yet we have not settled on a final strategy. We consider two approaches here. In terms of neurological function one obvious solution would be to set SSS = 0 should the patient die. This would give a negative change from baseline, which seems appropriate. A patient who enters the trial with an SSS of 20 and dies has a change from baseline of —20. However, a patient who arrives with SSS = 40 and at the end of the study has a score of 10 has a worse change from baseline (—30) and therefore a greater negative impact on the effect at the treatment assigned.

An alternative approach is to take the score at death to be the minimum SSS over the period of the trial, including baseline. A patient who improves but then dies nets a 0. A patient who declines by 10 points and then dies scores —10. But it might make more sense to count them the same. Indeed, the first patient may have passed through a change of

—10 while dying but happened not to be tested at that time. Again, survivors may have lower scores than those who do not survive.˚

Figure 8.7. Example of generated data including deaths. Upper left panel: death scored as zero. Upper right panel: death scored as minimum score observed in the trial. Deaths are indicated as circled pluses. Lower left panel: estimated dose-response curve for Curve 1 (flat dose-response curve) but including deaths.



The plots of the data (Figure 8.7, top two panels) show the distribution of deaths across doses. The first panel shows that scoring death as zero separates the distribution of scores for patients who die from that those who survive. The second panel (upper right) shows slightly more overlap by setting the score for death to the minimum SSS over time.

The assumed dose-response curve in this example is the flat dose-response curve (null Curve 1) and so the effect of including deaths in the data can easily be seen. For this reason we have also investigated dose-response curves that are not monotonic (Curve 7). Figure 8.6 shows results of our simulations (without deaths) in this curve. Doses assigned have adapted to the form of the dose-response curve and there are fewer patients allocated in the extremes of the dose range.

Appendices

Appendix 1a: Dose allocator initialization file

MAIN OPTIONS (DEFAULT IN [ ])

VERBOSE {0; 1; 2; 3}; specifies the level of output. 0 suppresses all output (except error messages), 3 includes extensive output (for debugging) [2]

ALLOC-RULE {1; 2; 3; 4}; the method used for computing the dose allocation;
1=decision theoretic myopic design, i.e., the optimal dose is computed by maximizing in expectation a utility function related to "learning" about the dose-response curve.
2=posterior quantiles: The posterior quantile rule allocates the next patient with probabilities p = 0:2; 0:1; 0:6; 0:1 at z = 0; z10 ; z50 and z90, respectively
3=random dose allocation. Let z = (z1 , ..., zJ) denote the list of allowable doses. Random dose allocation assigns with equal probabilities p = 1=(J + 1) one of the doses zj or placebo z = 0.
4=same as 3, but using the allocation probabilities given in dose-pz [1]

GEN-RULE {3; 4}. Same as alloc-rule, to be used by the generator [3]

DATA-DEATH {0; 1; 2}. Indicator for coding of death. See Section 9.5 for details. Three alternative methods are implemented. Minimum score: use the minimum observed (rescaled) score as final response (data—death=0); Last score carried forward: use the last observed (rescaled) score as final response (data—death=1); and Zero score: substitute yi = 0 as final score (data—death=2).

ALLOC-MIN__PLACEBO (0; 1), minimum allocation probability for placebo [0.1]

TERM-DYNAMIC {0; 1}, flag for using the dynamic terminator. Set the flag to 0 to get the static terminator. See Section 10.4 [0]

TERM-SIMULATE {0; 1}, flag for including forward simulation. Useful for experimenting with different weights for sampling cost and final "nugget" in the utility function. In normal operation the flag is set to 1 [1]

TERM-2NDSTUDY {0; 1}, flag for considering a second independent study in the definition of utilities for the terminator. See Section 10.2

RECO-R1RAND {1; 2; 3}, selects alternative methods for the recommender. Allocation proportional to expected utility (reco-R1rand 1), allocation strictly by maximum expected utility (reco-R1rand 2), or uniform allocation at any dose with a response within a given interval around the dose with maximum expected utility (reco-R1rand 3). Only relevant under alloc-rule 1 [1]

RECO-K {0; 1, ..., K — 1}, number of patients today before the next one (recommender only [0]

SIM-STOP {0; 1}. If sim-stop 1 is set then the simulator stops when either termination or switch to R2 is recommended. Otherwise (sim-stop 0) the simulator runs until the complete data set patients.data is read.

DIMENSIONS

DATA-NMAXR1, the maximum number of patients on the data file, i.e., the maximum number of patients in the learning phase R1 [500]

DATA-NMINR1, the minimum number of patients in R1 before switching to R2 [150]

DATA-NMAXR2, maximum number of patients in R2 (relevant only under the terminator). [500]

DATA-MAXNI, maximum number of observations per patient [14]

DATA-DIM, dimension of the design vector xi in the logistic regression. If relevant, this includes a constant for the intercept and a coefficient for dose [3]

DATA-NCOV, number of covariates (data-dim (excluding dose and intercept). The design vector has the additional intercept and could include interactions [1]

DATA-COVBAR, average covariate value (used to center the logistic regression [180.0]

DATA-SSS0-BAR, an average value for base SSS score. This does not need to be exact. The value is only used for plausibility checks and possibly for centering regressions [45]

INCLUSION CRITERIA: BOUNDS ON SSS AND TIME SINCE STROKE TO BE INCLUDED IN THE STUDY.

DATA-SSSLO, lower bound on SSS [10]

DATA-SSSHI, upper bound on SSS [50]

DATA-TSHI, upper bound on time since stroke (in minutes) [480]

ESTIMATOR (MCMC TUNING PARAMETERS):

MCMC-NITER, the number of iterations in the MCMC simulation to generate a posterior Monte Carlo sample [100]

MCMC-DISCARD, the length of an initial transient in the MCMC simulation which is discarded as "burn-in". [10]

MCMC-NBATCH, batch sizes in the posterior MCMC simulation. Simulations are only saved after every mcmc-nbath-th iteration [5]

MCMC-SPL__M {0; 1}, indicator for inference on the longitudinal model.
Note: If less than 50% of the patients have complete day 90 observations then mcmc-spl__m 1 will be automatically set.

MCMC-QZ, four probabilities ffk , k = 1, ..., 4 (as fraction [0; 1]). Posterior inference on the quantiles EDffk is reported in q.upd and q.bayes.

ALLOCATOR:

ALLOC-ALPHA [0; 1]. The allocator minimizes expected posterior variance for g() = zff , where ff =alloc-alpha. See Section 9.6 [0.50]

ALLOC-K, maximum number of future patients on the next day in the myopic rule. Computing time increases linearly with alloc-k [3]

ALLOC-K__S , maximum number of future patients on the next day in the myopic rule when running the program as simulator. Since in simulator mode the program calls the allocator currently only once a week, this needs to be larger than alloc-k [8]

YALLOC-NU, number of simulated experiments for dose z when evaluating expected utilities for the myopic rule [200]

ALLOC-NKNOTS, number of knots in the non-linear scatterplot smoother used to find the maximum expected utility in a scatterplot of simulated utilities against possible doses. [11]

ALLOC-J, number of allowable doses (compare alloc-z). [10]

ALLOC-COVS, covariates x0 of the "typical" patient used for computing the optimal dose allocation described in Section 9.6. [180.0]

ALLOC-Z, list of allowable doses zj.
[0.00 0.125 0.25 0.375 0.50 0.625 0.75 0.875 1.00 1.125 1.25 1.375 1.500]

DLM-DELTA (0:0; 1:0). Sets the discount factor ffi for the normal dynamic linear model (NDLM, see Section 9.2). A smaller discount factor leads to more smoothing, a larger discount factor leads to less smoothing. The setting should depend on the number J of allowable doses (alloc-J). For example, for J = 10, we recommend a discount factor between 0:10 and 0:30. The program allows a different discount factor for each dose.
[0.1 0.1 0.2 0.3 0.4 0.5 0.5 0.5 0.5 0.5 0.5 0.4 0.3 0.2 0.1 0.1]

Technical description: The discount factor ffi defines the evolution variance Wj in the NDLM as Wj = ffiCj—1 , where Cj-1 is the posterior variance-covariance matrix at the next lower dose, given the responses for doses j = 0, ..., j — 1.

DLM-G01, list of values G01 for the normal dynamic linear model (NDLM). Setting G01 equal 1.0 gives a 2nd order polynomial NDLM (i.e., fitting local line segments). Setting G01 equal 0.0 gives a 1st order polynomial NDLM (i.e., fitting local horizontal lines). When choosing this option, consider adjusting dlm-delta accordingly.

RECOMMENDER:

RECO-DUP, **RECO-DLO**, relative size of the steps up and down used for the additional randomization under alloc-rule 3 and reco-R1rand 3

TERMINATOR:

TERM-NM, dimensions of the discretized grid on m to define the table in (m; s) for the terminator (See Section 10 for a definition of (m; s)). [10]

TERM-NS, dimensions of the discretized grid on s to define the table in (m; s) for the terminator. [10]

YTERM-NSIM, Monte Carlo sample size for the number of simulated trials to evaluate u1. [100]

TERM-NU2, Monte Carlo sample size for the number of simulated trials to evaluate u2. [100]

TERM-PPWEEK, patients per week. [7]

TERM-DH*, cutoff for significance calculations. [0.0]

TERM-MSMIN, minimum difference (m1 — s1) to allow switch to R2 [0.5]

TERM-LIMITN, indicator for limiting the static terminator to simulations with sample size Nt, close to N [0]

TERM-C1, cost per patient. [10]

TERM-C2, payoff per point in advantage over placebo at optimal dose. [5000]

SIMULATOR:

SIM-NDISCARD, initial burn-in in MCMC runs for each day of the simulated clinical trial. [100]

SIM-NITER, MCMC run length for each run of the allocator in the simulated clinical trial. [6100]

SIM-FIRST__T, first run of the terminator after x weeks. [100]

SIM-DWEEK__T, run terminator after each x weeks. [100]

RANDOM NUMBER GENERATOR:

RV-SEED1 and **RV-SEED2**, seeds for the r.v. generator (long integer).
[981963 6843437]

PRIOR AND INITIALIZATION OF THE DOSE/RESPONSE CURVE MODEL

DR-B, initialization for b [—1 1.0]

DR-MB, prior mean b for b. [—1 1.0]

DR-SB, prior variance covariance matrix b for b.
[diag(10.0,10.0)]

DR-MY, shift in the logistic regression in the dose/response curve. [7]

DR-DY, scale dy of the logistic regression in (1). [7]

DR-M__DY, prior mean dr of dy. [10]

DR-S__DY, prior s.d. oedr of dy. [10]

DR-SIG2, initial value for the measurement error variance σ^2 in (1). [50.0]

DR-SIG2__0, fixed value for σ^2 for simulation of patients (in the simulator).
[50.0]

DR-M__SIG2, prior mean σ^2 for σ^2 . [50.0]

DR-NU__SIG2, degrees of freedom σ^2 for the inverse gamma prior on σ^2 . [10.0]

INITIAL VALUES AND PRIOR PARAMETERS FOR THE LONGITUDINAL DATA MODEL

LONG-M, initial values of the regression coefficients m1 and a1 in the longitudinal data model
[0.00 0.00 —1.90 0.90 —1.15 0.96 —1.64 0.98 —0.29 0.96 —0.72 0.94
—0.43 0.98 —0.53 0.99 0.16 0.93 —0.97 1.03 —0.23 0.99 —0.27 0.98 0 1]

LONG-MM, prior mean m for m1.
[0.00 0.00 —1.90 0.90 —1.15 0.96 —1.64 0.98 —0.29 0.96 —0.72 0.94
—0.43 0.98 —0.53 0.99 0.16 0.93 —0.97 1.03 —0.23 0.99 —0.27 0.98 0 1]

LONG-SM1, prior variance covariance matrix m for each row of m1
[diag(1.0, 0.01)]

LONG-S12, initial values for the variances s12.
[0.0 24.0 19.2 12.3 7.8 9.0 3.8 6.8 19.5 12.8 2.7 2.6 2.6]

LONG-MS1, prior mean s2 for the s2 vector.
[0.0 24.0 19.2 12.3 7.8 9.0 3.8 6.8 19.5 12.8 2.7 2.6 2.6]

LONG-NUS1, d.f. s2;j in the inverse gamma priors for each s2j.
[10 10 10 10 10 10 10 10 10 10 10 10 10]

Appendix 1b. The Initialization File init.gen for the Generator

GENERATOR:

GEN-NDAYS, number of days to cover in the generated data file [100]

GEN-PDAYS, probability distribution for the interarrival times T of subsequent patients: Pr(T = 0); Pr(T = 1); Pr(T = 2); Pr(T = 3) (does not need to be normalized).
[0.4 0.4 0.2 0.1]

GEN-LASTONLY {0; 1}, indicator for simulating only day 90 responses (usefull for debugging).
[0]

GEN-CONSTR {0; 1}, indicator for constraining simulated SSS scores to be between 0 and 58 (useful for debugging). [1]

GEN-F, for simulation mode only: expected response for each of the doses zj.
[10 10 10 10 11 12 13 14 15 16 17 18 18]

GEN-PZ, relevant only under alloc-rule=4: allocation probabilities for each dose in alloc-z.
[1 1 1 1 1 1 1 1 1 1 1 1 1]

GEN-PDEATH, probability of death, expressed as a logistic regression in dose z and baseline score SSS0. Let (j) = 1=(1 + exp [—j]) denote the logistic function. Then Pr(death|z; SSS0) = a0 + a1 z + a2 (SSS0 — SSS0) + a3 (SSS0 — SSS0)2 :
The option sets the four coefficients a0 a1 a2 a3. Use (—10000000) for Pr(death) = 0.
Use (—2:0000) for a uniform P(death|z; SSS0) = 12%.
[—10000 0 0 0]

GEN-PWEEK, probability of death occuring in week t, t = 0, ..., M , given that death occurs.
[0 1 1 1 0 0 0 0 0 0 0 0 0 0]

9. DISCUSSION

Drug development is an extremely expensive enterprise. Part of the expense is due to the sluggardly nature of modern clinical trials. A fixed (and generally) large number of patients are accrued in trials having deterministic designs. It is true that many trials employ optional stopping via boundaries such as those proposed by OÕBrien and Fleming (1979). But these designs are very conservative. The reason is that they must niggardly protect the overall Type I error rate of the frequentist tradition.

The Bayesian approach has no such inhibitions. But Bayesians must live in a world still dominated by frequentist attitudes. In this report we use the flexibility and naturally adaptive nature of the Bayesian approach but with an overall goal that is consistent with frequentist tradition. It is a Bayesian cake with frequentist icing.

A ubiquitous problem in drug development is that the correct dose of a drug may never be adequately established. Our design takes the dose-finding issue very seriously, while simultaneously deciding whether any dose of the drug is effective.

The first phase of our study uses Bayesian updating information about dose-response and uses a normal dynamic linear model. Should the results be sufficiently compelling (as judged by a decision analysis) and should a highly effective dose be identified then development of the drug enters a standard randomized confirmatory trial in which drug is compared with placebo. If the first phase establishes the drug as effective and if the confirmatory trial verifies its effectiveness then this combination would be very powerful evidence submitted to a regulatory agency in seeking marketing approval.

There are several virtues to our approach:

(1) It avoids wasting patient resources investigating uninteresting doses.

(2) It stops the trial as soon as it becomes sufficiently clear that the drug is not effective.

(3) It moves into confirmatory phase when the evidence suggests that such is appropriate.

(4) It adapts to the actual variability in the accumulating data, perhaps stopping earlier than would be predicted (small variance) and perhaps continuing longer than would be predicted (large variance).

(5) It uses a longitudinal model of the course of stroke in predicting final results for those patients having incomplete data, and learns about this model on the basis of the patients accrued to the trial.

The only disadvantage in our approach is that it requires a system for communicating with participating centers. This system must be able to update the statistical software based on recent information and it must be able to relay the doses assigned in very short order. Such a system is expensive, but its cost is tiny in comparison with the potential savings that accrue from using an efficient design.

10. References

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis, New York: Springer-Verlag.

Berry, D.A. (1993) A case for Bayesianism in clinical trials (with discussion), Statistics in Medicine, 12, 1377-1404.

Berry, D.A. and Stangl, D.K. (1996) Bayesian methods in health-related research. In Bayesian Biostatistics, (D.A. Berry and D.K. Stangl, eds.), New-York: Dekker, 3-66.

Carter, C. K., and Kohn, R. (1994), On Gibbs sampling for state space models, Biometrika, 81, 541-553.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review, Statistical Science, 10, 273-304.

DeGroot, M. (1970). Optimal Statistical Decisions. New York: McGraw Hill.

Fr¿hwirth-Schnatter, S. (1994), Data augmentation and dynamic linear models, Journal of Time Series Analysis, 15, 183-202.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (eds.) (1996). Markov Chain Monte Carlo in Practice, Chapman and Hall.

J¿rgensen, H.S., Nakayama, H., Raaschou, H.O., Vive-Larsen, J., St¿ier, M. and Olsen, T.S. (1994). Outcome and time course of recovery in stroke. Part I: Outcome. The Copenhagen Stroke Study. Archives Physical and Medical Rehabilitation,76,399-405.

Lindenstr¿m, E., Boysen, G, Christiansen, L.W., Rogvi Hansen, B. and Nielsen, B.W. (1991). Reliability of Scandinavian Stroke Scale. Cerebrovascular Disease, 1,103-107.

O'Brien, P.C., Fleming, T.R. (1979). A multiple testing procedure for clinical trials, Biometrics, 35:549-556.

Scandinavian Stroke Study Group (1985). Multicenter trial of hemodilution in ischemic stroke. Background and study protocol. Stroke,16,885-890.

Thall, P.F., Simon, R.M., and Estey, E.H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes, Statistics in Medicine, 14, 357-379.

Thall, P.F., and Russell, K.E. (199x). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. to appear in Biometrics.

Tierney, L., (1994), Markov chains for exploring posterior distributions, Annals of Statistics, 22, 1701-1762, (With discussion).

Verdinelli, I. (1992). Advances in Bayesian experimental design. In Bayesian Statistics IV, (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds), pp. 467-48, Oxford: Oxford University Press.

West, M. and Harrison, P.J. (1997). *Bayesian Forecasting and Dynamic Models.* Second Edition. New York: Springer-Verlag.

Whitehead, J. and Brunier, H. (1995). Bayesian decision procedures for dose determining experiments, Statistics in Medicine, 14, 885-893.

Whitehead, J., and Williamson, D. (1998). Bayesian decision procedures based on logistic regression models for dose-finding studies, Journal of Biopharmaceutical Statistics, 8, 445-467.