Why Do So Many Phase 3 Clinical Trials Fail?

Part 1: The Effect of Deficient Phase 2 Trials in Therapeutic Areas with High Failure Rates in Phase 3 Studies

By: Anastassios D. Retzios, Ph.D.

Contents

Address correspondence to the author at: Bay Clinical R&D Services, 2417 Canyon Lakes Drive, San Ramon, California 94582 -- Address email to: aretzios@gmail.com

Copyright © 2009 Bay Clinical R&D Services - All Rights Reserved

A. Summary

The development of new drugs in oncology and in stroke, both leading causes of mortality and disability is exceedingly important both for the pharmaceutical industry and the general public. It should be a public health priority. However, the number of Phase 3 clinical trials that fail in these critical areas is very high, raising the costs of development and delaying or canceling the introduction of new and more effective therapies.

Although on the surface it may appear that the challenges in oncology and stroke drug development are very different, in fact, the causal factors that lead to failure in Phase 3 trials are very much the same. A major contributor to the high failure rate is inadequate Phase 2 programs that provide sub-optimal information for the "go/no go" decision to move to Phase 3 and the design of the Phase 3 trials.

Deficient Phase 2 programs are either inadequately designed, do not contain the full complement of studies, incorporate endpoints that provide limited or misleading information regarding the efficacy of the test agent, or are improperly executed. The specific challenges vary with the therapeutic area. In oncology, trial designs and endpoints utilized for cytotoxic compounds may not be appropriate in the development of the newer targeted, cytostatic therapeutic agents. In the treatment of stroke, the designs and the endpoints typically employed so far may not have the sensitivity and reliability to allow investigators to define the effects of neuroprotective compounds,

In this article, some of the problems plaguing the Phase 2 programs in oncology and stroke are summarized and certain proposed solutions are presented and evaluated. Despite the rather narrow focus of this article in terms of therapeutic areas, the fault lines in Phase 2 studies have a universal dimension in clinical research.

B. Introduction

For many clinical development professionals, the failure of a pivotal trial to achieve its primary endpoint is a very difficult personal experience. The development of therapeutic agents is a lengthy process and it absorbs a substantial proportion of one's professional life. Therefore, other situations notwithstanding, failure at this stage is emotionally wrenching. It also has an ethical dimension. Large numbers of

patients (occasionally thousands) have been exposed to a compound that did not provide a possibility for clinical improvement. Clinical research within the context of ethical, well-controlled and mostly randomized clinical studies is a rather new discipline that emerged after the end of the 2nd World War. It involves a structured approach to developing chemical compounds and biologics after these entities have shown promising therapeutic potential in *vitro* and in animal testing. Clinical trials in early phases of development (Phases 1 and 2) explore their safety and potential clinical benefit in humans. If these studies are well implemented, they will weed out novel drugs and biologics with serious safety issues and with questionable activity and also collect adequate information for the design of the Phase 3 program. Only agents that have demonstrated acceptable safety and efficacy progress to the pivotal phase of development (Phase 3) where they must demonstrate their efficacy within the regulatory requirements of definitive proof to gain marketing approval.

1. Attrition Rates in Pharmaceutical Research

This is, at least, the theory. The outlined approached should limit failures in pivotal studies if the Phase 2 program is well implemented. Unfortunately, this is not the case. The rate of failure in pivotal studies is still quite substantial, standing recently at about 45%.[1] In certain key areas and with more novel compounds, the failure rate has been substantially higher. For example, for biopharmaceuticals that entered clinical trials in oncology throughout the 90's, the success rate was a very low 13%.[2] Recent estimates by the FDA have lowered this estimate to approximately 8%.[3] European-based industry groups are in agreement with these estimates.[4] If one looks at all agents developed for oncology applications in the same time frame, only 5% of those that entered clinical development ever reached approval; approximately 60% of those that had apparently successful Phase 2 programs failed in Phase 3 studies.[1] For certain CNS applications, especially neuroprotection in ischemic stroke or head trauma, the attrition rate is probably the highest of any field. Well over fifty compounds have been tested in numerous clinical trials but none was proven clinically beneficial.[5,6] The industry seems to be falling below the numbers required for replacement of commercially successful compounds required to maintain revenues. The rarity of new drugs and biologics is fueling a number of

mergers and acquisitions in the industry. which is failing to provide adequate solutions in many devastating diseases.

The high rate of failure rate in late stages of development is placing a substantial burden on the pharmaceutical industry. Costs are rising because failures occur later in development and quite often in Phase 3.1 Thus, getting adequate information for a "go/no go" decision at least in Phase 2 would substantially reduce development costs and may allow more compounds to be tested. In addition, improvements in design that may allow studies to obtain rigorous data in a shorter period for time and enhancements in operational aspects of Phase 3 studies have the capability of reducing costs and minimizing the number of failures.

2. Why examine the causes of failure in drug development in cancer and ischemic stroke?

To examine the causes that lead to pivotal program failure and to attempt to find some solutions, we would be examining in some detail relevant issues in oncology and neuroprotection in ischemic stroke, areas in which development is plagued by high rates of failure. Cancer and ischemic stroke are the 2nd and 3rd leading causes of mortality in the US and stroke is the primary cause of disability. Thus, increasing the pace and success rate of development in these areas is of primary importance not only to the pharmaceutical industry and to clinical research professionals but to the general public as well. This not to say that other key therapeutic areas are not seriously plagued by critical and difficult to solve issues; they surely are. Although certain of the observations here may have a universal applicability, we will integrate eventually a discussion on sepsis, pain and certain CNS disorders such as Alzheimer's disease, in an expanded edition of these articles.

C. Searching For Answers

Defining the root causes of attrition of experimental drugs and biologics is the essential first step in an attempt to remediate the problem. However, examining clinical trial publications for causes of failure presents a number of challenges. In the absence of certain compelling findings, studies with "negative" findings are typically not published or are published after a substantial delay.[7] Even when

studies are published, crucial elements of the study such as statistical design are often missing, as surveys of the literature indicate.[8,9] It is also quite likely that in a published negative study, the information included may not necessarily reflect the original provisions under which the study was performed; usually, it is a retrospectively-defined analysis meant to tease out information for future studies or new directions in research. Thus, a substantial amount of knowledge that can be derived from development failures remains unavailable or has been compromised. To highlight the discrepancy between positive and negative study publications, a recent survey of reported breast cancer Phase 2 studies found that 80% had positive outcome.[10] Since regulatory authority surveys show that the percentage of successful Phase 2 studies in this indication is lower than 50%,[11] one can easily surmise the extent of under-reporting. Since all clinical studies for new medical entities (NMEs) performed under an IND (as well as federally funded ones) have to be reported to the database maintained by ClinTrials.gov as a requirement of the FDA Modernization Act (1997), this database can be potentially used to determine the true extent of under-reporting, a task as yet not undertaken by regulators. The problem is serious enough for editors of prestigious medical journals to recently promise higher publication priority to failed studies under certain conditions.[12]

Despite the problems of under-reporting, investigations of the causes of failure have attracted some attention. This is especially true in areas with a high number of failures such as neuroprotection in ischemic stroke and cancer. It is typical for researchers in any field to seek new study designs and new endpoints after a substantial number of studies of promising drugs/biologics have failed. The core problems in clinical development in these indications as well as others, many of which I had previous involvement with, will be discussed below and a number of proposed solutions will be presented and evaluated.

D. Phase 2 Development Issues

But let's start with the basics. A Phase 3 program would commence only if there is "positive" information in prior development. Logically, the reasons that pivotal studies may fail despite the early promising results can be divided into two main areas: (a) misleading information collected in Phase 2, that the test compound is efficacious in the targeted indication and selected population, while the reverse is true; and (b) problems with the design and implementation of the Phase 3 program.

In this article, we shall concentrate on the reasons why Phase 2 programs provide either misleading or inadequate information. Problems resulting in inaccurate information from the Phase 2 program can be numerous. Before examining the main ones, we need to review the position of a Phase 2 program within the overall scheme of drug development.

Typical development, as stated above, commences with Phase 1 programs. These programs consist of a number of studies usually performed in healthy volunteers. They obtain safety and pharmacokinetic information and define, if appropriate, the highest tolerable dose of the compound or biologic that can be administered. Drugs that present unacceptably high toxicity profile in this stage are discontinued. "Go/no go" decisions on these grounds at the end of Phase 1 are relatively easy.

Phase 2 programs are comprised typically of randomized and controlled studies in the patient population of interest. They attempt to define if a pharmacological activity can be discerned by a number of objective or subjective assessments (depending on the indication) and how well this activity compares to that of a placebo or an active control (usually, the currently available treatment). They also attempt to determine the appropriate dosing of the drug, to expand safety information, to define drug-drug and drug-food interactions and to collect certain information pertinent to the design of a Phase 3 program. Safety remains a strong component of Phase 2 programs. It is certainly possible for serious toxicities to be observed for the first time in a Phase 2 trial, as certain drugs present an unacceptable toxic profile only in disease states. This has been the case for certain NMDA antagonists tested in neuroprotection; they exhibited acceptable toxicity in phase 1, but were shown to have increased mortality in Phase 2.[13]

At the conclusion of a successful Phase 2 program, the development team would have established, within certain predefined bounds, that that the candidate drug has "activity" in the indication tested and the target population of interest. A well-run Phase 2 program would also have provided information about the appropriate dose for the pivotal studies and provided an estimate for the sample size required for the Phase 3 trials. The team must then decide if the drug meets the "desirability quotient" for further development taking into account institutional constraints and policies, as well as input from regulatory authorities. A very important "go/no go" decision usually hinges on that information at the conclusion of this program.

The high rate of failure of Phase 3 studies,[1] especially in certain indications, demonstrates that many Phase 2 programs are inadequate in making such determinations. Surveys that examined the rate of success of Phase 3 studies after "positive" Phase 2 data highlight this point. Zia *et al.*[14] surveyed published reports of Phase 3 studies in cancer treatment from 1998 to 2003, and matched them to their equivalent Phase 2 studies. Forty-three such reports were identified and linked to 49 "positive" Phase 2 studies. Of the Phase 3 studies, only 28% resulted in positive outcomes, despite the "encouraging" Phase 2 results. In another recent survey of pivotal studies of targeted agents in cancer treatment from 1985 to 2005, Chan *et al.*[15] showed that only about 50% of Phase 3 trials in cancer with positive Phase 2 results were successful.

In a methodological approach, we can divide deficiencies in the Phase 2 clinical trials in the following categories: (a) inadequate design; (b) endpoints with a tenuous connection to clinical-benefit-based Phase 3 endpoints; and (c) improper execution. These categories are not mutually exclusive; a failed program may span a number of them.

1. Designs of Phase 2 Clinical Programs, Potential Issues And Impact On The Selection Process

Inadequacies in design can cover a wide variety of clinical study elements: sample size, endpoints, randomization appropriate controls, schedule of assessments, and a variety of other elements appropriate for the correct execution of a complete Phase 2 program. Some of the deficiencies are omissions, others result from the "state of the art" at the time of study conception and others are imposed by specific circumstances such as funding and corporate priorities. It is not unusual for a Phase 2 program to suffer in an attempt to shorten timelines to approval. In my experience, Phase 2 programs are the development area that comes most often under scrutiny for potential trimming in order to shorten timelines. The problem has become more glaring as development has moved into more complex disease states and into indications with established treatments in place. Proceeding with inadequate Phase 2 programs under these conditions only enhances the possibility of failure in Phase 3.

More often than not, Phase 2 programs are incomplete. For example, rather frequently in oncology, dose-ranging studies are not performed as part of the Phase 2 program despite the fact that they are strongly encouraged by regulatory authorities.[16] The maximum tolerable dose (MTD) is defined in Phase 1 and then the maxim of "more is best" usually prevails. Although such an approach has been more or less effective with cytotoxic compounds, newer targeted compounds that have little toxicity present different challenges. In this case, dose ranging studies must be incorporated into the Phase 1 and Phase 2 program. As toxicity cannot any longer be the main determinant for the dose, the test compound effect on the tumor target must be the primary endpoint of dose ranging studies. For anti-angiogenic compounds, effects on the vascular density of the tumor and related parameters may constitute useful endpoints.[17]

Sample sizes are smaller in Phase 2 programs, as at this phase one attempts to be efficient in screening out compounds for further development. Many methods are utilized to keep patient numbers small while providing reliable information for a "go/no go" decision. For example, surrogate*/pharmacodynamic/ biomarker endpoints in which test compounds are expected to exhibit a larger difference from the control than the "clinical-benefit" based endpoints are routinely employed. These surrogate endpoints are usually proximal or distal pharmacodynamic effects. In addition, designs that are "statistically efficient" are utilized, although several of them result in under-powering studies and obtaining dubious results. Underpowered Phase 2 studies that result in "positive" outcomes may suffer from type 1 error. In a type 1 error, the null hypothesis (that the drug is not better than the control) is rejected and a "false positive" occurs. Despite the heightened probabilities for false positives and when other elements of design are optimal, powering Phase 2 studies to a one-sided alpha of 0.10 -or even 0.20- is regarded acceptable for certain studies;[18] however, it is important to always consider the inherent implications of the

* The term "surrogate" here is has a flexible application. A true surrogate would substitute fully for the clinical-benefit endpoint. However, there are a variety of endpoints that are not pharmacodynamic- or biomarker-based, they are clinical in nature but do not correspond fully to the clinical benefit endpoints typically used in pivotal studies. E.g., Progression-free Survival (PFS) or Time to Progression (TTP) are such endpoints. These endpoints are discussed in greater detail in Section D.2.

statistical compromise. In certain indications such as oncology, designs utilizing small number of patients have been developed to provide reliable assessments - despite the small sample size-. The large number of oncology studies and the relative scarcity of subjects make such designs an imperative. Neuroprotection in ischemic stroke is an area in which innovation in design may allow a far larger number of compounds to be tested by limiting sample sizes. This would be the shot in the arm that this field of research badly needs. It is important to emphasize, however, that reconciling small sample sizes and robust information in Phase 2 is always an intricate balancing act that requires both well-argued statistical methodology but also the establishment of culture of objectivity and certain detachment.

a. Design of Phase 2 Clinical Studies in Oncology: A Field in Evolution

Since oncology is a therapeutic area with a substantial rate of failure, it is important to examine the conditions that render Phase 2 data in this therapeutic area less likely to be adequate for selection of compounds for further development.

Oncology has always been rather unique because of the challenge of the disease to both patients and treaters and the perceived need to move to pivotal studies what was originally regarded as a small number of new chemical entities with "carefully" balanced toxicity/benefit ratios. There is a number of unique designs for Phase 2 studies in this therapeutic area, some reflecting a certain "gestalt" in the field of keeping sample sizes small and moving as many promising therapeutic agents and/or regimens to Phase 3 as quickly as possible.

Until recently, most chemical entities in cancer treatment were cytotoxic and were investigated by the most ubiquitous and certainly unique design for Phase 2 studies in this indication, the single-arm, two-stage clinical trial.[19] The basic rationale for such a design remains quite simple and rests on two assumptions (a) tumors are unlikely to regress without pharmacological intervention (although certain tumor types show high spontaneous regression rates); and (b) the percentage of response for the standard treatment (which constitutes a historical control) can be adequately defined . Tumor regression measured by standardized criteria is usually the measure

of "efficacy" in these studies although "polynomial" endpoints combining tumor shrinkage and measures of disease progression have been proposed and implemented.[20] If the "percentage of effectiveness," (usually set at 20% above the assumed effectiveness of the historical control, with a rejection error at 10%) is met in stage 1 with a very small cohort, a larger cohort is enrolled in stage 2; the sample size in stage 2 depends again at where one sets the "percentage of effectiveness" and the standard error. Using optimal designs, these studies can be powered for 80-95% and at a significance level of 0.05 with relatively small number of patients (35-55).[21,22] Three-stage designs have also been proposed[23] but remain rare because they are complex to implement. Of course, positive results in such clinical trials and the corresponding decisions to proceed to Phase 3 depend heavily of how well the historical controls have been defined.

Outside oncology, the use of historical controls is rather frowned upon and usually discouraged by regulatory guidance.[24] And for a good reason. Historical controls may not be the appropriate comparator for data collected prospectively, because of differences in concomitant treatment, demographics, study entry criteria, the time and type of assessments, the methodology of measurement and a number of other study provisions.

On the basis of the criticisms outlined above, it is apparent that the existence of a well-defined and appropriate historical control is of primary importance for the utilization of non-randomized designs. But how well have these historical controls been defined and how does their quality influence the "go/no go" decisions? Vickers *et al*.[25] examined the robustness of historical controls in Phase 2 studies published in two prominent oncology journals, the "Journal of Clinical Oncology" and "Cancer," between 2002 and 2005. Of the examined studies, 52% utilized historical controls. On the basis of specific criteria, the definition of the response rate of the historical control was regarded as appropriate, not appropriate or none. Phase 2 studies that defined their historical control with appropriate methodology were statistically less likely to result in positive results and to declare that the examined compound worthy of further testing. Thus, one may conclude that the use of non-randomized and staged designs with historical controls should

be regarded acceptable only if these historical controls are rigorously defined and sampling errors are taken into consideration.

Another popular choice for Phase 2 studies in oncology is the randomized selection clinical trial. In this type of a design, a number of treatments are randomized but no control is utilized and the drug (or drug combination) with the largest percentage of efficacy progresses to further development.[26,27] Such studies have a 90% power to detect the most efficacious treatment assuming a 15% absolute increase in response rate. No particular concern is paid to a type 1 error (false positive). This design, although probably more robust than the single-arm, two-stage design and capable of differentiating well between several competing treatments, suffers from the same weakness as a filter for a "go/no go" decision to Phase 3. The decision to proceed to Phase 3 depends ultimately on a comparison to a historical control.

To overcome the problems of these designs and to obtain more reliable data for a "go/no-go" decision, multi-arm, randomized designs with concurrent standard treatment or placebo (if possible) controls have been suggested.[28,29,30] The reasons that randomized designs have become more attractive recently is because historical controls are inadequate or non-existent for the newer cytostatic agents that do not result in tumor shrinkage but may have a substantial impact on overall survival (OS).[31] These cytostatic agents may prohibit tumor growth and metastases but do not result in substantial reduction in size during the shortened period of observation of Phase 2 studies. Since OS takes a number of years of observation to be established, randomized Phase 2 studies with cytostatic agents normally utilize disease progression endpoints (Section 2.D.a.).

Randomized designs may be the most appropriate for the evaluation of cytostatic agents; however, they have a number of disadvantages. Patients may be reticent to enter a study that would result in an assignment to either placebo or standard treatment. In addition, "the statistical efficiency" of these designs increases the possibility of false positives. Liu *et al.*[32] criticized such randomized and controlled designs as capable of discerning only extraordinarily high differences in treatment because of their low power and the possibility of being misused for treatment decisions. The authors

calculated that the "false positive" error rate in these studies may be as high as 40%, substantially higher than what one may expect even with an alpha as high as 0.20. Can these designs become more robust? Taylor *et al.*[33] simulated a randomized two-arm study and a non-randomized one with the same number of subjects. The authors assumed certain uncertainty in the definition of the historical control and an additional uncertainty about the historical control perception by the investigators. Running the simulation shows that increasing the sample size from 30 to 80 patients increases the probability of a correct decision (to proceed to Phase 3) more with a randomized design than with a single-arm study.

Although moderately larger sample sizes may increase the dependability of randomized studies to provide accurate data, patient heterogeneity in these still small samples remains a major challenge.[34] If certain covariates, relevant to the potential outcome (such as patient molecular phenotype) are not balanced, it is quite possible that a positive or negative difference from control may be due to an imbalance in these covariates. In large Phase 3 studies, the randomization of large numbers of patients in study arms usually results in balanced groups, but in smaller Phase 2 trials the presence of randomization alone may not result in an appropriate balance. Of course, patient heterogeneity does not only affect Phase 2 randomized designs, but it does compromise their main argument of being able to provide a more reliable comparison of the efficacy of the test compound to the current treatment. Certain Phase 2 studies are certainly taking this issue under consideration, especially for cytostatic compounds with specific targets. For example, in a study of weekly docetaxel plus trastuzumab versus weekly paclitaxel plus trastuzumab in non-small cell lung carcinoma, the patients randomized to the arms of the study were also stratified on the basis of HER-2 protein expression.[35] This was a reasonable approach since trastuzumab targets the HER-2/neu receptor. The study did not reveal any advantages for these treatments, but the approach was valid. Thall and Wathen introduced recently two Bayesian-based hierarchical designs that address the issue of patient heterogeneity head on and allow treatment to focus on subtypes that respond to the treatment.[36]

Randomized discontinuation designs have also been introduced to deal with several of the disadvantages of the randomized trials. They have been utilized in a wide variety of indications prior to being advocated for oncology development. Essentially, these designs enrich the study for patients of a given post-treatment characteristic (e.g., rate of tumor growth). The resulting more homogeneous population reduces variance and increases the statistical power of study. These designs also overcome the reticence of patients in enrolling in purely randomized designs. In such a design, the trial is performed in two stages.[37] In the first stage, all patients are treated with the test compound; those with stable disease are then randomized to either the test compound or the control (placebo or current treatment) and disease progression is assessed. Accession in stage 1 ceases when the stage 2 randomized sample size is achieved. Of course, the dangers of such a design for a "go/no go" decision are substantial; the patient population has been enriched and the drug effect has been amplified. In addition, there may be difficulties in the implementation of such designs. Blinding in stage 2 may be compromised if the active treatment presents a definitive toxic profile.

It is apparent from the above that the specific design chosen much reflect the nature of the compound tested and the degree of certainty in prior data, if a historical control may be contemplated. Even for cytostatic agents, tumor regression endpoints may provide useful clues as to the desirability of further development as El-Maraghi and Eisenhaurer recently showed.[38] One may tend to gravitate towards actively controlled, randomized designs that are more typical in most areas of clinical development, but one should also remain aware that limited subject numbers within a heterogeneous patient population and "statistical efficiency" may not result necessarily in results more dependable than those of the single arm designs.

We do not have dependable metrics as to which Phase 2 designs are likely to provide more robust information. A baseline may be provided by the survey of Zia *et al.*[14] In matching "positive" Phase 2 studies with their equivalent Phase 3 trials for a five year period (1998-2003), he estimated a success rate of 28%. In this survey, all but two of the positive Phase 2 studies had a single-arm, non-randomized design. It remains an open question if the randomized Phase 2 designs can show a higher level of success in Phase than the 28% of

the single-arm, non-randomized ones. A future survey may address this question.

A number of researchers strongly advocate the abandonment of non-randomized designs[39] while others maintain that they have a place in development in oncology, especially in cases in which there is strong consensus on response rates of historical controls and the study population is well characterized.[40] Although, personally, I would always tend to gravitate towards controlled and randomized designs, I recognize that there are countervailing considerations: randomized and controlled studies have a lower level of appeal to cancer patients, and, this factor affects enrollment adversely. In addition, it is easier to terminate studies based on non-randomized and staged designs as they are open-label studies and Bayesian analysis of responses at a certain time point in stage 1 may provide a good indication of the futility (or not) of continuing the study.[21] Thus, if there is a well-defined historical control that meets stringent criteria, the possibility of a non-randomized, staged study should be examined. Patient heterogeneity (various covariates and molecular phenotype) must also be taken into account either as part of the study design or data analysis in order to increase data robustness and to indicate sub-populations more likely to respond to treatment. This may limit the target population of a Phase 3 study –as well as the indications of resulting drug label- but this is a challenge that each organization must face within its institutional constraints.

In response to some of these considerations, adaptive designs have also been recently introduced to the study of oncology, promising a better selection of active regimens and doses and adequate power to discern a treatment effect. We will be discussing adaptive designs not only in oncology but in other indications in a forthcoming article on this site.

b. Study Design In Neuroprotection: Consistent Failure Breeds Uncertainty and a Few New Ideas.

Despite some encouraging results in some serious neurological disorders in the early part of the 2nd half of the 20th century, progress in this field has been very slow. Stroke, a leading cause of mortality and disability,[41] is a therapeutic area in which pharmacological treatments have made very little

inroad in providing clinically beneficial solutions. Reperfusion agents such as alteplase, a recombinant tissue plasminogen activator (rt-PA), have been approved although the path to registration has been difficult.[42] Other stroke treatments, mainly compounds attempting to salvage ischemic tissue have not been as successful. Despite a substantial number of potential treatment agents and numerous trials, not a single putatively cytoprotective compound has shown clinical benefit. Recent reviews summarizing the development efforts in the field by Labiche and Grotta[43] and Gingsberg[6] outline the failed efforts and highlight the daunting challenges in this area.

What accounts for this high failure rate? There can only be three reasons: (a) ischemic tissue in the aftermath of a stroke cannot be salvaged and would transform to an infarct irrespective of any intervention; (b) the compounds tested were inadequate to achieve neuroprotection; and (c) the design and endpoints of clinical studies were deficient to detect a beneficial effect.

Experimental work that laid the foundations for neuroprotection, the experience with the reperfusion agents and some recent studies with non-pharmacological approaches indicate that reason (a) is likely not true. There were a number of industry-academia efforts to minimize reason (b) by specifying the preclinical information needed prior to any clinical trial, but these have met with modest success, as we will be discussing later on. Undoubtedly, the purported mechanism as well as the preclinical testing behind several of the compounds pursued remains suspect. However, the examination of the record reveals that reason (c), in conjunction with the regulatory framework, is most likely the main culprit for the rate of failure in this field.

If in oncology one has the challenge to select an appropriate design for early clinical studies and to deal with the uncertainty of their statistical compromises, the issues in neuroprotection study design are far more elemental. The near total lack of success in the clinic has caused researchers to progressively question almost every element of design.[44] Although the overwhelming majority of the Phase 2 studies have been randomized and controlled and this approach appears to be almost universally accepted, most other issues of design remain the focus of debate. Overall, study quality is

low. In 2001, Kidwell *et al*.[5] reviewed the state of ischemic stroke trials (both Phase 2 and Phase 3) and scored the quality of clinical studies on a scale from 0 to 100 on five pre-specified quality criteria. The resulting analysis showed that clinical studies in this indication were sub-optimal in design despite progressive increase in quality in the decade prior to the report.

At the end of the 20th century, the unremitting failure rate with neuroprotective compounds in stroke led to the creation of the Stroke Therapy Academic Round Table (STRAIR), an industry-academia joint effort to rectify matters. A certain belief prevailed at that time, that the reasons for failure were not really those of clinical designs; they were due mostly to inadequate compounds entering clinical testing. STAIR was supposed to remedy the situation with clear guidelines as to which compounds should enter clinical trials.[45] How effective these criteria were remains a point of contention. There is no evidence that the compounds that entered clinical research after the publication of the STAIR recommendations adhered strictly to these recommendations. This was the conclusion of Philip *et al.* in a recent review.[46] Although the field saw a lot of promise in NXY-059, a free radical trapping compound, reviews after its failure in a pivotal study (SAINT II) concluded that its preclinical development did not adhere to the STAIR guidelines.[47]

STAIR placed some emphasis in optimizing elements of clinical study design to assure success, although certain of these guidelines were, at best, vague,.[48] A variety of additional guidelines were issued in the ensuing years by the same group.[49,50,51] Despite the proposed guidelines and their supposed implementation, a number of major studies in the first decade of the 21st century also failed, most notably the SAINT studies of NXY-059.[52,53] The first study with NXY-059, SAINT I, detected an improvement in one of the primary endpoints (the modified Rankin scale).[52] This led to a larger Phase 3 trial, the SAINT II, which did not show any clinical benefit.[53] Subsequent re-evaluation of the SAINT I statistical analysis showed that the purported improvement in the modified Rankin scale was more illusory than real and likely resulting –among other causes- from not adjusting for multiplicity of testing.[54]

So, despite almost a decade following the STAIR guidelines, the field is in search for a clinical design that would be able to detect what some now regard as a rather modest treatment effect of neuroprotection. Study inclusion criteria are still a focus of discussion. The whole theory behind neuroprotection presupposes the presence of an ischemic penumbra which, supposedly, can be prevented into turning into an infarct by pharmacological and non-pharmacological interventions. The presence of the ischemic penumbra can be demonstrated by functional neuroimaging[55,56] although arguments regarding the best possible methodology for its detection are ongoing.[57] Unfortunately, very few studies so far have enrolled patients with a verified presence of an ischemic penumbra; no study has established the required extent of this penumbra for a potentially successful pharmacological intervention. The extent of the penumbra in the area affected by stroke varies substantially, if the limited study by Heiss *et al.* [58] provides representative data. In that study, the penumbra contributed 8 – 34% of the final infarct volume whereas 51 to 92% of the final infarct volume was unsalvageable within 3 hours after stroke. Such a discrepancy in salvageable tissue among patients and the inability of the studies to compensate for it, is likely a contributory factor to the failure of the effort so far. The recent failure of the DIAS-2,[59] a clinical trial in which only stroke patients with a verified presence of an ischemic penumbra were enrolled, should not indicate that such an approach is not valid. In the DIAS studies, desmoteplase, a plasmin activator, was utilized to provide recanalization in stroke patients 3 to 9 hours after presentation. The selected patients had an MRI-defined ischemic penumbra. The study failed to show improvement over placebo despite "encouraging" results of in the Phase 2 studies (DIAS-1 and DEDAS).[60,61] In fact, these studies are a prime example of advocacy prevailing over relatively scarce and contradictory data and illustrate vividly what one needs to avoid in a "go/no go" decision process for initiating a Phase 3 study. Both DIAS-1 and DEDAS were very small studies (102 and 37 subjects, respectively) and their results regarding reperfusion with desmoteplase were discordant. The reasons for the failure of the DIAS and DEDAS studies may have been agent-specific. The DEFUSE study, in which alteplase (rt-PA) was utilized at 3 to 6 hours after stroke, showed that patients with a verified presence of a

penumbra, were statistically more likely to have a favorable outcome following reperfusion.[62] These studies have helped substantially to advance the state of the art of ischemic penumbra detection and the topography and evolution of lesions in stroke.[63] Hopefully, these advances would be utilized in the development of newer agents.

Another issue that confounds the results of smaller, Phase 2 studies in neuroprotection is stroke severity. Usually, stroke severity is assessed by the National Institutes of Health Stroke Scale (more information about the NIHSS can be found in Section D.2.b). Stroke severity is a strong predictor of outcome. Unfortunately, most often than not, there is no attempt to stratify patients on severity during randomization, despite the serious implications on the validity of the data. The interpretation of both the GAIN I[64] and the AbESTT I[65] Phase 2 data was confounded by the imbalances in stroke severity between groups. In addition, including too many patients with mild to moderate stroke reduces assay sensitivity as these patients are most likely to make a full recovery without intervention.

In addition, neuroprotective drugs should not be expected to have any effect in hemorrhagic strokes that constitute approximately 10% of all strokes. Despite this fact, only certain recent studies have excluded patients with CT-based evidence of hemorrhage and this was mainly due to the fact that the compound tested increased hemorrhagic risk.[66]

Nor have studies differentiated the treatment population on the basis of the location of the stroke. The great majority of tested compounds in clinical studies address synaptic ischemic events (in grey matter) and not strokes that occur in white matter although ischemic events after stroke may be distributed in both grey and white matter. Ho *et al.*[67] examined 77 patients who participated in neuroprotection clinical studies and determined that 95% of those had white matter involvement in the infarct volume and, on average, white matter made up about 50% of the total infarct volume. Thus, a substantial number of patients were treated with compounds that were not designed to address their pathological condition and, thus, any potential efficacy signal was too diluted to have been detectable. In summary, provisions for patient selection in neuroprotection studies so far has

contributed into including in the studies a highly heterogeneous group of patients, many of whom would have been regarded unsuitable for treatment if the appropriate parameters were carefully assessed. This factor may have had a cardinal impact in the failure of these studies to show any clinical benefit.

Another major issue regarding patient inclusion in clinical trials has been (and remains) the allowable time window for intervention. Most of the clinical studies in the previous decade did not place any serious restrictions to time to treatment, although all preclinical work indicated that treatment as early as possible was necessary for a favorable outcome.[68] There is a general agreement now to limit enrollment to patients that can be treated within the first 6 hours although intervention within the first 3 hours is highly desirable. In clinical trials performed during the last ten years, studies such as GAIN, AbESTT and SAINT, intervention was limited to 6 hours after occurrence, but, again, only a minority of patients was treated within 3 hours of occurrence.

The duration of drug administration is also uncertain, but this parameter would probably need to be individualized for each agent tested. Again, the absence of any success has led to many questions regarding the length of intervention and to advocacy for longer treatment periods.[69] In most trials, pharmacological intervention is limited to the first 24 hours although it has been established for some time now that neuronal death continues for days after the ischemic event.[70] For some agents, longer periods of intervention may not be possible because of their toxicity or because they may impede full recovery.[71] The fact that very few Phase 2 studies were performed to examine optimization of dosing prior to proceeding to Phase 3 is quite indicative of the need for flexible designs that would assess a variety of parameters with a limited number of patients prior to a "go/no go" decision. Overall, there has been a tendency in this field to move on to pivotal studies without many clear answers in Phase 2. As noted above, in the AbESTT I study, the investigators noted that the results were ambiguous and most likely due to an imbalance in stroke severity between the placebo and test groups.[65] However, despite these issues, a larger Phase 3 study, AbESTT II was initiated but it enrollment was halted at midpoint due to safety concerns (hemorrhage) and lack of clinical

benefit.[66] The very same problems were also reported by the GAIN investigators after the analysis of data from the gavestinel Phase 2 clinical studies.[64] Again, larger Phase 3 studies were initiated despite the lack of a clear indication of efficacy.

Another problem for clinical trials in stroke is the lack of agreement on the optimal treatment modality when attempting a pharmacological intervention to achieve neuroprotection. It has been argued that neuroprotection should be administered in conjunction with thrombolytic treatment as the best way of allowing the tested compound to reach the ischemic tissue.[72] However, only patients that can be treated within 3 hours of stroke onset and have no CT-evidence of hemorrhage are eligible for treatment with alteplase (rt-PA), making difficult –albeit not impossible- to maintain an acceptable patient accession pace if the enrollment criteria match those of alteplase use. Virtually all recent studies have utilized a 6-hour treatment window and only AbESTT excluded patients with CT-based evidence of hemorrhage.[65,66] Subgroup analysis for thrombolytic treatment in neuroprotective studies has not revealed any evidence of augmentation of efficacy, but most of these analyses were not adequately powered.[73] In fact, the use of alteplase in most of the recent studies in patients that qualified for its administration may have confounded any evidence of efficacy for the neuroprotective agents, because the overall treatment algorithm among patients was not standardized. However, excluding alteplase-treatable patients presents a serious dilemma to the design and implementation of trials since they likely constitute the best population for testing neuroprotective compounds.

Designs to quickly select compounds for further development have not been much in evidence in neuroprotection. Most Phase 2 studies incorporated the typical clinical endpoints consisting of widely used impairment and disability scales, in multi-arm randomization (usually treatment vs. placebo) and were powered for the typical significance level (0.05) and power (>80%). Thus, even Phase 2 studies in this indication tended to be quite large and operationally demanding. For example, the SAINT I trial treated approximately 1772 subjects and it was actually underpowered for the multiple endpoints used.[54] It seems intuitive, therefore, that designs must be devised that would allow a quick but accurate screening of compounds and

select the ones for further development. Towards this end, Palesh *et al.*[74] proposed a non-randomized, single arm design for futility Phase 2 studies, an adaptation of the non-randomized optimal designs in oncology. Of course, such a design depends on an accurate determination of the efficacy of treatment of the historical control (π_0). The authors evaluated their design in previous reperfusion studies in stroke† and concluded that the design would have selected the compounds for further development at a fraction of the subjects treated in the studies examined. For example, applying the design to the failed ATLANTIS study (Alteplase® in stroke), only 169 patients would have been required to declare futility instead of the 613 actually enrolled prior to discontinuation of the study (its intended sample size was 968).

The MRI Collaborative Group[75] also introduced a design based on MRI imaging-based endpoint, the infarct expansion ratio (IER, the ratio between the final infarct and the initial ischemic tissue volume). The introduction of this endpoint was based on a study of clinical outcomes and MRI imaging of 259 patients in a variety of institutions and countries. The authors determined that the selected endpoint, the IER, had a strong correlation with the clinical outcome measured on the modified Rankin scale, the endpoint of choice in Phase III clinical studies in neuroprotection. On the basis of the observed differences, the authors believed that a study targeting a 20% increase in therapeutic effectiveness, 80% power and a two-sided $\alpha = 0.05$ would require 99 patients per arm for a dichotomous outcome (IER$\leq$ 1.0 and IER> 1.0) and 61 patients for a continuous one. Such a design would certainly allow a faster screening of promising compounds and a better chance of success in Phase 3 studies (which, so far, has proven elusive). The utilization of imaging technologies in endpoints of studies in neuroprotection is more extensively discussed in Section D.2.b.

Mandava *et al.*[76] also attempted to streamline the compound selection process for pivotal studies by producing a "3-dimentional" model utilizing data from a variety of existing studies to structure a "control" population. The model

† Reperfusion treatment with fibrinolytic enzymes has been so far the only pharmacological intervention in stroke that has been found effective despite a number of failed studies

includes age and NIHSS scores as baseline variables and mortality and modified Rankin Scale score of 0 – 2‡ as outcomes. Compounds that achieve an efficacy level in the test arm higher than the upper 95% confidence boundary of the model control would be regarded as meriting further examination. The authors applied this methodology to certain studies investigating non-pharmacological interventions (lasers and ultrasound) and determined that Phase 3 studies for these interventions are warranted. They also applied the model to the SAINT I and AbESTT I studies and the results showed that neither NXY-059 nor abciximab merited further examination (at least as utilized in the Phase 2 studies). The authors concluded that this model can be utilized to screen compounds in smaller, non-randomized studies

In conclusion, what the area of neuroprotection needs right now is a reliable methodology that would allow for a speedier screening of compounds in Phase 2 than has been possible before and provide adequate information in the decision process to move to the Phase 3 stage. These designs and more suitable endpoints (discussed in Section D.2.b) may allow the pharmaceutical industry to restart development in this area.

2. Endpoints of Phase 2 Studies, Correspondence to Clinical Benefit and Impact on Design

I decided to address endpoints separately from other elements of design simply because their impact is extensive and because they constitute a unique point of failure. In Phase 3 studies, the primary endpoint must incorporate a well-defined and generally accepted clinical benefit. In earlier development, one can utilize surrogate endpoints (usually pharmacodynamic parameters or disease progression criteria) that may not correspond directly to measurable clinical benefit but provide strong evidence of pharmacological activity. Surrogate endpoints utilized in Phase 2 trials present specific advantages in this phase of

‡ In the modified Rankin scale (mRS), 0 = no symptoms, 1= no significant disability despite symptoms and 2 = slight disability but able to look after own affairs without assistance. Scores 3 – 5 denote higher level of disability and the score of 6 is assigned to death.

development. They can be assessed after a relatively short period of follow-up or/and they may result in more pronounced differences between treatment groups than those possible with the clinically-relevant endpoint. Thus, they can provide for a speedier and less costly selection process, as the studies can be shorter in time and utilize fewer patients than studies based on clinically relevant endpoints. In many cases, these endpoints they may allow numerous Phase 2 studies to be performed in order to assess drug-drug interactions and dose ranging for the collection of appropriate information prior to the onset of the pivotal program. The use of surrogate endpoints is definitely desirable in areas in which the clinically relevant endpoint requires lengthy observation; degenerative central nervous system diseases or oncology are examples of such therapeutic areas. In multiple sclerosis, the clinically relevant endpoint consists of structured observations with validated instruments over a long period of time (usually 2 years) that documents the presence or absence of progressive disability.[77] Obviously, such endpoints in Phase 2 would substantially delay the selection process. Phase 2 programs are usually based on the effect of the test agent on the size and number of brain lesions within a period of few months. In oncology, overall survival (OS) is the main clinical endpoint, but the period required in assessing is lengthy. Thus, the utilization of alternate endpoints in Phase 2 studies such as tumor regression or progression-free survival becomes imperative, despite controversies on their use.

a. The Endpoint Conundrum in Phase 2 studies in Oncology

Discussion on endpoints in Phase 2 trials in oncology is inextricably linked to the debate on study design. For the longest time, the test compound effect on tumor size (response/no response based on a percentage of tumor reduction) was the "gold standard" of endpoints in this phase of development. Standardized methodology for assessing tumor response was formulated to standardize assessment and has progressively evolved over the years. Tumor response and patient assignment to response categories is performed by the utilization of standardized assessment tools such as the WHO criteria and RECIST.[78,79] RECIST, which entered usage in 2000, was developed to address certain deficiencies of WHO and has been recently updated to RECIST 1.1.[80] On the basis of these instruments, patients are characterized as

full or partial responders, exhibiting no change or as having progressive disease. For the majority of solid tumors and for cytotoxic compounds, tumor response correlates well with overall survival and the possibility of regulatory approval, although in certain cancers such as melanoma and renal cell cancer, no such correlation exists.[81] In addition, spontaneous regression rates complicate assessment as these may be relatively high in certain tumors,[82] the reason why differences in response rates between the historical control and the active treatment (π-π_0) in non-randomized, multi-staged designs are usually gated at ≥ 20%. Tumor response, despite criticisms, remains the most utilized endpoint in Phase 2 studies. The surveys of literature by El Maraghi and Eisenhauer[38] and Vickers *et al*.[25] found that at least 70% of the Phase 2 studies in oncology utilize this endpoint.

Recently, the advent of cytostatic agents has challenged the utilization of tumor response, as well as its assessment tools, as an endpoint in Phase 2 development. Targeted, non-cytotoxic but cytostatic agents may stabilize the disease without substantially reducing tumor size; alternatively, tumor regression may be either slower or less pronounced to be assessed satisfactorily by either WHO or RECIST. It has thus been argued that the application of the RECIST criteria creates an artificial dichotomy (responders, non-responders) based on the percentage of shrinkage that may or may not correspond to clinical benefit. Thus, proposals have been advanced to rely on tumor size as a continuous variable rather than a dichotomous one.[83,84] Obviously, such an approach would be highly dependent on the frequency of assessments. Also, because historical controls may be severely lacking in continuous tumor size data, its application would require randomized, multi-arm designs.

Additional endpoints have also been proposed that address the progression of the disease, such as progression-free survival (PFS) and/or time to progression (TTP).[85,86] These may be more appropriate for cytostatic agents although neither of these endpoints is particularly predictive of overall survival, which remains the primary endpoint of pivotal studies.[87]. In the Chan *et al*.[15] survey mentioned above, which investigated success in Phase 3 studies after favorable Phase 2 data, the use of PFS and TTP showed a trend

to be associated with success at the pivotal stage, although this trend was not statistically significant.

Despite the appeal of PFS and TTP for the development of cytostatic drugs, there is a wide agreement summarized by Dhani *et al.*,[88] that these endpoints present certain disadvantages that need to be compensated by adjustments to the design of the clinical trials. They are highly influenced by the frequency of assessments. In addition, disease progression for a standardized period of observation (a few months to a year, depending on the rate of progression of the tumor type under examination) is quite variable among patients. Therefore, it is difficult to assign lack of progression to pharmacological action. In addition, both PFS and TTP are very susceptible to investigator bias. Obviously, randomized and controlled designs are strongly recommended if PFS and TTP are selected as the primary endpoints of the Phase 2 trial.[31] PFS and TTP may be problematical if utilized in non-randomized, multi-staged studies. Historical controls may be deficient and, depending on the period of observation for PFS and TTP, moving to the 2nd stage of these single-arm studies may necessitate interruption of the study. Certain researchers view this delay as an advantage rather as a disadvantage, allowing for some over-enrollment and careful examination of the data.[30] However, the extended period of data collection has the definitive drawback of increasing dropouts and an attendant impact on the validity of a trial based on such a binary and time-dependent endpoint. In any case, surveys of data of previous clinical trials in specific cancers has provided historical controls of certain robustness and do allow PFS to be utilized in non-randomized designs.[85,89] In the Vickers *et al*.[25] survey of Phase 2 trials using historical controls, 24% of studies utilized disease progression criteria, including PFS.

Since PFS and TTP have a number of disadvantages, and since tumor regression assessed by RECIST may not be the best tool for cytostatic agents, newer imaging techniques may provide a better methodology for evaluating objective response. These techniques included fluorodeoxyglucose positron emission tomography (FDG-PET), contrast-enhanced magnetic resonance imaging and spectroscopy. As we look to the future, they may have a definitive role in the assessment of the effects of both a cytotoxic and cytostatic compounds on tumors in Phase 2 studies.[90]

To ascertain the validity of the newly proposed endpoints for Phase 2 studies such as PFS and TTP, El-Maraghi and Eisenhauer[38] examined 89 Phase 2 studies of 19 targeted agents and showed that cytostatic compounds that did not exhibit any objective response in tumor regression did not gain regulatory approval, although they may have displayed substantial non-progression rates as assessed by PFS. On the basis of these results, the authors argued that non-randomized designs utilizing tumor regression endpoints – which accounted for 70% of the examined reports in this survey - are still capable of providing adequate information for a "go/no go" decision. It should be pointed out, however, that the same report identified cytostatic agents that gained regulatory approval with tumor regression rates well below the typically set rate of 20% (in several cases, the rate was <10%). Obviously, non-randomized designs targeting such efficacy rates would require larger sample sizes than has been the norm, as Booth *et al.*[40] realized.

Biomarkers can certainly be introduced as secondary endpoints in Phase 2 studies.[91,92] These biomarkers can ascertain if the appropriate target of the cytostatic compound is affected or/and they can define the molecular phenotype characteristics of patients that respond (or do not) to treatment. Such information may be quite important in designing follow-up clinical trials. The development and validation of these tests has not been robust thus far, but their utilization would allow for a better selection of individuals that may benefit from treatment, thus enhancing the statistical power and the possibility of success of Phase 3 studies. Utilization of biomarkers may lead to substantial fragmentation of indications but it may lead to more pronounced clinical benefits and an enhanced clarity for the use of newer targeted therapies. As the search for newer and better endpoints becomes more and more necessary to better select efficacious compounds, one needs to counterbalance this effort5 with the always-wise approach of keeping studies relatively simple and not overly demanding for patients. Discouraging enrollment has its own attendant pitfalls.

b. Neuroprotection Endpoints: The Problem with Disability and Outcome Scales

Clinical trials in neuroprotection originally utilized loosely constructed and non-validated endpoints. However, for the last twenty years, clinical trials in this indication have incorporated one or more validated impairment or global function scales as elements of the primary endpoint. The STAIR recommendations also focused on the utilization of impairment or global function scales to show clinical benefit, although they suggested the use of imaging techniques to discern the extent of activity of the drug. Among the most utilized of these scales are the National Institutes of Health Stroke Scale (NIHSS), the Barthel Index (BI), the modified Rankin scale (mRS) and the Glasgow Outcome Score (GOS).

A few comments on these scales would be appropriate in understanding their function and their limitations in neuroprotective research:

- The NIHSS assesses various neurological functions assigning scores (from 0 up to 4) to each one.[93] Ascending scores indicate increased loss of function. The NIHSS correlates moderately to strongly with infarct volume as determined by CT and MRI (r = 0.4-0.8).[93]
- The Barthel Index (BI) is a scale that assesses the stroke patient's capability for self-care and mobility.[94] It has a moderate correlation to infarct volume (r = 0.3 – 0.5).[95] The score for normal in this scale is 100 with lower scores denoting increased disability. This scale is strongly predictive of the long-term outcome,[96] although it suffers from a "ceiling effect" as symptoms related to cognition, language, sight, emotional status and pain are not included.[97]
- The modified Rankin scale (mRS) is in common use to assess disability after stroke.[98] It is a 7-point ordinal scale. Usually, a one point shift on this scale is regarded as clinically significant because the categories included are rather broad. It has a moderate correlation to the infarct volume (r = 0.4 – 0.5).[99] In clinical research, mRS is regarded desirable because it allows for smaller sample sizes in trials. Statistical considerations and analysis are based usually on a dichotomized outcome: Scores of 0 to 2 are regarded as successful outcome and 3 to 6 as an

intervention failure.[100] Lai and Duncan maintained that utilizing its scale as a continuous endpoint would enhance its utility.[101]

- The Glasgow Outcome Scale (GOS) is a 6-point ordinal scale, very similar to the mRS.[102] It is widely used in the clinical setting, although in clinical research it is not typically utilized as a primary endpoint for power considerations. It emphasizes physical disability and does not define social and functional deficits as well as the mRS.[103]

How sensitive are these scales? Some of them exhibit little sensitivity to changes. Dromerick *et al.*[97] examined the sensitivity of mRS and BI, among other less common scales, in 95 stroke patients. The mRS detected changes in 55 and the BI in 71 of these patients. Young *et al.*,[104] in simulations using the GAIN studies as a point of reference determined that the NIHSS is more sensitive than either BI or mRS and if dichotomized at ≤ 1 it would allow a substantial reduction in sample size in clinical studies if used as the primary endpoint. However, dichotomizing the scales hardly provides any decisive advantage in obtaining reliable outcomes. Sulter *et al.*[105] found out that the dichotomization of the mRS and BI had been used inconsistently and that it had been easier to define an unfavorable rather than a favorable outcome. Also, dichotomizing the scales in an arbitrary fashion affects the interpretation of the data. In the ECASS II study (the alteplase study in Europe), when the definition of favorable outcome changed from mRS ≤ 1 to mRS ≤2, the results of the study achieved statistical significance.[105,106]

The utilization of these scales in multicenter trials also raises issues regarding the consistency and reliability of assessments by the raters involved in these studies and the overall effect of rating consistency in the results of the studies. For such a core element of the primary endpoint, the inter-rater consistency effects on outcomes have not been widely (if at all) investigated. A small number of studies have been performed and results varied for each of the scales. BI showed a very high rate of consistency between raters, probably because of its in-build redundancy.[107] So did the NIHSS, but only after training programs and certifications within the bounds of a clinical study.[108,109] When a recent survey examined the NIHSS ratings of thousands of clinicians, the results were less than stellar and the authors remarked that

repeated training had little effect and such rating inconsistency can affect the results of studies.[110] The mRS is susceptible to substantial inter-rater variability.[111,112] In the absence of a structured interview process, inter-rater agreement was very poor (43%) and it improved to only 81% after the utilization of a structured interview.[113,114] Thus, in any project in which these scales are used, there is a need for substantial and ongoing training that has to overcome bias and habits developed over a long period of practice. Even if successful, an additional source of variance is introduced that may be influencing outcomes in smaller Phase 2 studies.

The NINDS trial, which investigated the role of alteplase in stroke, was one of the few positive studies in stroke treatment.[115] Notably, it utilized the typical rating scales in a novel way that has not been attempted since. In this trial, the NIHSS, mRS, BI and GOS scales were combined into a single global outcome endpoint on the assumption that a treatment effect would be detectable in all scales. It was shown that this approach increased statistical power (each individual scale as an endpoint would have failed to detect a positive effect) and it also avoided the Bonferroni adjustment for multiple primary endpoints (of dividing the alpha by number of tests).[116]

The brief description of the scales and their correlation with the infarct volume highlight the problem that one faces in developing drugs for neuroprotection. The drugs are designed to limit infarct volume but their efficacy in clinical studies is evaluated on the basis of endpoints that assess disability, which is influenced by other factors besides infarct volume. It seems far more reasonable, at least for Phase 2 studies to utilize the only relevant endpoint that makes sense, reduction in infarct volume. This was the position was advocated a decade ago by Saver *et al*.[117] The investigators, who utilized their results of the study of tirilizad in stroke (RANTTAS),[118] noted a moderate correlation between CT-derived infarct volume and impairment scales such as BI, NIHSS and GOS. The limitations of the CT methodology may have partially impaired the data, as infarct volume was not determined in approximately one third of patients. Newer technology allows a better definition of the infarct volume.[119] Thus, Savitz and Fisher also reiterated the argument in favor of centering at least Phase 2 studies on reduction of infarct volume.[120]

The utilization of reduction of infarct volume as an endpoint in studies in neuroprotection is augmented by the fact that compounds that were eventually unsuccessful in exhibiting clinical benefit were also incapable of reducing infarct volume. Van der Worp *et al.* [121] did not see evidence of infarct volume reduction with tirilizad, a compound tested extensively in acute stroke without exhibiting clinical benefit.[122] Also, Warach *et al.*[123] did not detect any change in infarct volume with gavestinel, the NMDA receptor glycine-site antagonist investigated in the negative GAIN studies. Clinical studies with magnesium in stroke patients also failed to show any clinical benefit and Kidwell *et al.*[124] were unable to detect any reduction in infarct growth in a sub-study. A recent study of the use of alteplase in stroke between 3 and 6 hours after admission (EPITHET), a non-statistically significant reduction in infarct growth was detected between baseline and Day 90 when compared to placebo.[125] However, his study was relatively small (101 randomized patients) and the treatment window for alteplase (3 to 6 hours post-stroke) has not been shown to be definitely efficacious.

However, one must examine this issue in the context of the pathological condition being addressed. Disability resulting from stroke is not always related to infarct volume but also to the location of the stroke in the brain: a very small infarct in a critical location may result in a more pronounced disability than a sizeable infarct in a less crucial area. This is the main reason for the moderate correlation of infarct volume with impairment as defined by global function scales. However, the recent efforts by Menezes *et al.*[126] may point the way in combining infarct volume with infarct location. These investigators have constructed brain atlases that assign a neurological deficit severity value to brain voxels (a 3-D defined pixel). In their study, the combination of infarct volume and location value from these atlases substantially increased the correlation to the NIHSS score ($r=0.79$, $P=0.032$) than infarct volume alone. These atlases are far from complete, but one can certainly see the potential of their utility as endpoints in clinical studies in stroke.

In summary, examining the primary endpoints utilized so far, one can certainly perceive that neuroprotective drugs, including reperfusion agents, have been struggling against substantial obstacles when assessed for activity

in clinical studies utilizing the typical scales. Even if they manage to limit infarct volume in a statistically meaningful manner, they may not be able to show corresponding changes in disability or global function because of the location of the stroke and the resulting moderate correlation with global function scales, as well as scale insensitivity, inter-rater inconsistency or scale ceiling effects.

Despite the recognition of this problem and the calls for change, many perceive that the regulatory framework which requires "clinical benefit" to be proven (within the defined statistical requirements) prior to approval is, at this time, not amenable to change. Indeed, the EMEA, specifically states that "MRI measures cannot be considered recognized efficacy surrogate endpoints and they may only supplement – but cannot replace- proper clinical efficacy criteria, at least in Phase 3" in its 2001 considerations for development in acute stroke.[127] Similar pronouncements are likely by the FDA. The FDA does allow "accelerated approvals" on the basis of surrogate endpoints (provided it accepts them as such) but it demands additional studies to verify "clinical benefit". However, without a change in the regulatory framework for endpoints in this area, it is questionable if the impairment and global function scales can be relied to assess on their own effective pharmacological interventions. It is in this context that the Menezes *et al.*[126] atlases and their strong correlation with the NIHSS and the central repository of imaging data and tools outlined in the "Advanced Neuroimaging for Acute Stroke Treatment" meeting on September 7 and 8, 2007 in Washington DC[128] are vitally important. They may provide the foundations for better endpoints in stroke studies. In a recent editorial in the "Stroke" journal, Wardlaw sounded a cautionary note about surrogate endpoints such as MRI imaging of infarct volume.[129] The author noted that in the EPITHET study many patients died before being evaluated, thus confounding the results. This is indeed true, but surrogate endpoints of Phase 2 studies can be used in combination with other outcomes and scales that would compensate for their disadvantages. It may make sense to utilize composite endpoints in Phase 2 studies that include infarct volume at 30 or 90 days and/or location-based weighted severity with infarct volume scores, mortality and a combination of impairment scales. The elements of this endpoint may be differentially weighted in accordance with

sensitivity. Such composite endpoints would facilitate development of drugs that both reduce infarct volume but do not result in unwarranted toxicities and inhibit rehabilitation. Obviously, a number of studies would be required to carefully define and validate such endpoints, but in the absence of such progress, a very crucial therapeutic area would see little progress in the coming years.

3. Inadequately Executed Phase 2 Studies

Independent of design and endpoint considerations, many studies fail to collect appropriate information in Phase 2 programs because of inadequate execution. In a lot of cases, the quality of the data is poor because of numerous protocol violations and deviations. Good clinical practices (GCPs) failures are multiple and infect virtually all studies. Any study in which protocol violations[§] are noted in ≥ 20% of enrolled subjects is likely unreliable. Multicenter clinical trials in which a large number of professionals is involved are very susceptible to GCP failures in the absence of rigorous central management, frequents audits, intensive monitoring, and pervasive and ongoing training. Dispersed organizations and the presence of participating centers in a variety of countries certainly amplify the possibilities of GCP breaches.

Beyond GCP violations, "go/no go" decisions are strongly affected by the desire to succeed. Many Phase 2 studies attempt to provide "proof of concept" in highly controlled situations with a small number of investigators and very homogeneous populations, although the demographics, prognosis and concomitant treatment of these patients is unlikely to be encountered in the pivotal studies. Thus, the specific treatment algorithms in a small number of centers may provide certain support for "efficacy" but the data collected would prove woefully inadequate for the design of a Phase 3 study. The result is that the therapeutic effect is overestimated and thus, the sample size of the pivotal study is underestimated. Consequently, the pivotal studies would fail to achieve the desired endpoint. In certain cases, when the endpoint is supported by

§ Protocol violations usually include inappropriate patient inclusion and non-protocol specified treatment among others and are be prospectively declared in the study statistical analysis plan.

examination of test results that may be subject to investigator bias, such as radiological data, the lack of an independent central review may result in responses much higher than an independent assessment board would assign.[130,131] In the Zia *et al.*[14] analysis of oncology studies discussed above, it was shown that Phase 2 studies that led to mostly negative Phase 3 results relied mostly on investigator analysis of radiologic data, whereas the Phase 3 trials mandated a central review.

4. Beyond Design and Endpoints: Funding and Resources

Of course, inappropriate designs, inadequate endpoints and faulty execution of the Phase 2 program are only part of the explanation of the high rate of failure of Phase 3 studies. The "go/no go decision" for commencing a pivotal program after the conclusion of Phase 2 is contaminated by other considerations. These are mostly institutional, organizational and funding issues.

In most situations, funding is intermittent to small and medium-sized pharmaceutical and biotech companies. It is certainly targeted to a short term outcome. There is substantial incompatibility between projections of earnings of venture or other investment capital managers and research, especially in the pharmaceutical industry. While one typically understands that research is not a linear process, income projections by funds make no such assumptions. Instruments are, of course, in place to seek additional funds, but this is a very time- and attention-consuming process which puts substantial pressure on the R&D team. Thus, the companies with inadequate earnings funded by the investment community rarely have the latitude or the time to utilize their technology in the development of a number of drug candidates in order to select the most promising ones after various rigorous Phase 2 programs. They are forced, by the level of their capitalization and their corporate boards, to "focus" on a very small number of therapeutic agents; more often than not, on only one. Under these conditions, it is sometimes human nature to see what one wants to see in a Phase 2 study, although the data may not be that clear. Several examples above illustrate this. The lack of funding also has a pernicious effect on the capacity to execute a program adequately. Money saving efforts, dispersed outsourcing and a variety of contractors makes it difficult to provide consistent training and adequate supervision; thus, the quality of data suffers.

Larger pharmaceutical companies are not immune of these considerations. R&D funding is always under pressure. In addition, in larger companies, organizational issues occasionally loom larger than funding. There are a number of "stakeholders" strongly invested in a particular compound who will spin data in various ways until a definitive failure makes any further maneuvering impossible. Strong association and identification with the project and the project team, as well as worries about the effects of a project discontinuation create an atmosphere in which it is very difficult to get a dispassionate review of Phase 2 data. "Gate" reviews of projects by committees of non-project affiliated personnel may help restore some objectivity.

E. Conclusions

It is quite evident that the challenges in designing and implementing a robust Phase 2 program are many. The main effort of that program would be to provide adequate information for an informed "go/no go" decision. Unfortunately, as we outlined above, Phase 2 programs are compromised for a variety of reasons and the subsequent Phase 3 trials fail to show a clinical benefit.

In certain indications, in which competition for eligible patients is substantial, keeping the Phase 2 clinical trials small requires an essential compromise between completing the program in a reasonable period of time and obtaining accurate information. How much compromise is acceptable has to be decided after a thorough review of the existing information.

A process that would be less prone to errors would involve a careful evaluation of the prior development and all historical data, if any, of the test compound or the compound class; the utilization of a design that fits well with existing information and provides an acceptable compromise between speed of development and obtaining reliable information; careful accounting of the heterogeneity of the Phase 2 study population both in trial design and data analysis; and the selection of an endpoint that is both suitable for the compound tested and corresponds well to the likely clinically beneficial endpoint that would be utilized in Phase 3.

It is very important in the planning of the Phase 2 program, to have a clear idea as to what needs to be achieved in the pivotal phase (Phase 3) both in terms of the population to be treated and the therapeutic advantage to be sought. In designing

development strategies for new drugs, a typical process would include the survey of the competitive environment and the construction of a target label. In many cases, best-case and base-case target labels can be constructed to more clearly delineate the pivots of the "go/no go" decisions at the end of each phase of development. In fact, it is highly encouraged, even by regulatory agencies, to design the overall development process with the target label in mind (most likely the base case). In this context, the Phase 2 program fits in a logical continuum and has an organic connection to previous and future development. Such an organic connection revolves around clearly understood "go/no go criteria" and makes the design of the program far easier and clearly understood throughout the whole development team.

As in most cases –and especially in clinical research-, the devil is in the details. Thus, protocol designs need to clearly written and communicated to all stake holders. The statistical sections should be fully developed both in terms of the assumptions for the design of the study but also of the data analyses to be undertaken. All elements of the protocol such as inclusion/exclusion criteria, randomization and stratification (if any), treatment modalities, schedule of assessments, compliance mechanisms, statistical considerations and data analysis should all be cross-checked against the target label to make certain that they comply strictly with the overall thrust of development. Although speed is required and innovative designs can provide assistance, skipping essential trials such as detailed pharmacokinetics, pharmacokinetics in subgroups of interest, dose-ranging studies, drug-drug/food interaction studies should not be sacrificed without a good understanding of their contribution to overall development. Abbreviated Phase 2 programs impose substantial risks on the pivotal phase.

Innovation is paramount. In clinical research, the pool of available patients for participation in clinical studies is becoming smaller and smaller as a multitude of compounds enter development. Expansion of clinical research to Asia, Eastern Europe, Latin America and Africa may provide a short respite with a lot of attendant complications. New study designs and endpoints that would allow the gathering of accurate information from smaller subject cohorts should be developed and their risks and benefits clearly outlined. Thus, making Phase 2 more robust is very crucial. Challenges to regulatory dogma, well-fortified and well-argued with appropriate scientific information are important to move innovative designs and endpoints forward to limit failures in subsequent Phase 3 studies. If we fail to do so,

several key areas of development may become "dead zones" despite the need for intense development both from the viewpoint of the pharmaceutical industry and that of public health. It is just possible that the current pace of development of new drugs may progressively slow even more than it is now. It is unlikely to see substantial development funding in areas in which the challenges to success have been and continue to be enormous and in which failures in pivotal programs are numerous.

F. Keywords

Phase 3 - Phase III - Phase 2 – Phase II - Clinical Trial – Failure Rate – Clinical Trial Design – Endpoints – Cancer - Oncology – Cytostatic - RECIST – Single-Arm - Randomized – Randomized Discontinuation – Tumor Regression – Progression-Free Survival – PFS – Time to Progresion – TTP - Stroke – Neuroprotection – Reperfusion – NIHSS – Barthel Index – Modified Rankin Scale – Glasgow Outcome Score – Ischemic Penumbra – Neuroimaging – MRI – Good Clinical Practices – GCP – Protocol Violations – "Go/No Go" Decision

G. Acknowledgements

I am very much indebted to William Cooper, MS, biostatistician for the Beardsworth Consulting Group, Inc., for his critical reading of this manuscript, his comments and suggestions several of which have been incorporated in the text.

H. References

1 Kola I and Landis J: Can the pharmaceutical industry reduce attrition rates. Nat Rev 3: 711-715, 2004

2 Pavlou AK and Reichert JM: Recombinant Protein Therapeutics- Success rates, market trends and values to 2010. Nat Biotechnol 22: 1513-1519, 2004

3 FDA: Challenges and Opportunities Report – March 2004. March 2004

4 Industry Success Rates 2004, Centre for Medicines Research International Ltd. CMR04-234R, May 2004

5 Kidwell CS, Liebeskind DS, Starkman S *et al.*: Trends in acute stroke trials through the 20th century. Stroke 32: 1349-1359, 2001

6 Ginsberg MD: Neuroprotection for ischemic stroke: Past, present and future. Neuropharmacology 5: 363-389, 2008

7 Krzyzanowska MK, Pinfile M and Tannock IF: Factors associated with failure to publish large randomized trials. JAMA 290:495-501, 2003

8 Mariani L and Marubini E: Content and quality of currently published phase II cancer trials. J Clin Oncol 8: 429-436, 2000

9 Thezenas S, Duffour J, Culine S *et al.*: Five-year change in statistical designs of phase II trials published in leading cancer journals. Eur J Cancer 40: 1244-1249, 2004

10 Perone F, Di Maio M, de Maio E *et al.*: Statistical design in phase II clinical trials and its implication in breast cancer. Lancet 4: 305-311, 2003

11 DiMasi, JA, Hansen RW, and Grabowski HG.. The Price of Innovation: New Estimates of Drug Development Costs. J Health Econom 22: 151–85, 2003

12 Cannistra SA: Phase II trials in Journal of Clinical Oncology. J Clin Oncol 27: 3073-3076, 2009

13 Hoyle L, Barber PA, Buchan AM *et al.*: The rise and fall of NMDA Antagonists for Ischemic Stroke. Curr Mol Med 4: 131-136, 2004

14 Zia IM, Siu LL, Pond GR *et al.*: Comparison of outcomes of Phase 2 studies and subsequent randomized controls studies using identical chemotherapeutic regimens. J Clin Oncol 23: 6982-6991, 2005

15 Chan JK, Ueda SM, Sugiyama VE *et al.*: Analysis of phase II studies on targeted agents and subsequent phase III trials. What are the predictors to success? J Clin Oncol 26: 1511-1518, 2008

16 International Conference on Harmonization (ICH)-Harmonized Tripartite Guideline: Dose-Response Information to Support Drug Registration. 10 Marh 1994

17 Eder JP and Zuckerman LS "Endpoints for the determination of efficacy of antiangiogenic agents in clinical trials" in "Cancer Drug and Discovery-Antiangiogenic Agents in Cancer Therapy, 2nd Edition" Eds Teicher BA and Ellis LM, Humana Press, Totowa, NJ, pp 509-524, 2007

18 Rubinstein L, Crowley J, Ivy P, LeBlanc M, and Sargent D: Randomized phase II designs. Clin Cancer Res 15: 1883-1890m 2009

19 Geehan EA: The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapy agent. J Chron Dis 13; 346-353, 1961

20 Dent S, Zee B, Dancey J *et al.*: Application of new multinomial phase II stopping rule using response and early progression. J Clin Oncol 19: 785-791, 2001

21 Simon R: Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10: 1-10, 1989

22 Schlesselman JJ and Reis IM: Phase II clinical trials in oncology: Strengths and limitations of two-stage designs. Cancer Invest 24: 404-412, 2006

23 Chen TT: Optimal three-stage design for phase II clinical trials. Stat Med 16: 2701-2711, 1997

24 FDA: E10 Guidance to Industry- Choice of Control Group and Related Issues in Clinical Trials. May 2001

25 Vickers AJ, Ballen V, Scher HI: Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. Clin Cancer Res 13: 972-976, 2007

26 Simon R, Thall PF and Ellenberg SS: New designs for the selection of treatments to be tested in randomized clinical trials. Stat Med 13: 417-129, 1994

27 Sargent DJ and Goldberg RM: A flexible design for multiple armed screening trials. Stat Med 20: 1051-1060, 2001

28 Rubinstein LV, Korn EL, Freidlin B, Hansberger S, Ivy SP and Smith MA: Design issues of randomized phase II trials and proposals for phase II screening trials. J Clin Oncol 23:7199-7206, 2005

29 Ratain MJ, and Sargent DJ: Optimizing the design of phase II oncology trials: the importance of randomization. Eur J Cancer 45:275-280, 2009

30 Korn EL, Arbuck SG, Pluda JM *et al.*: Clinical trial designs for cytostatic agents: are new approaches needed? J Clin Oncol 19: 265-272, 2001

31 Michaelis LC and Ratain MJ: Measuring response in a post RECIST world: from black and white to shades of grey. Nat Rev 6: 409-414, 2006

[32] Liu PY, LeBlanc M, Desai M: False positive rates of randomized Phase 2 designs. Cont Clin Trials 20: 343-352 (1999)

[33] Taylor JMG, Braun TM and Zhiguo L: Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm Phase 2 design

[34] Tuma RS: Examining heterogeneity in phase II trial designs may improve success in Phase III. J Natl Cancer Inst 100: 164-166, 2008

[35] Krug LM, Miller VA, Patel J *et al.*: Randomized phase II study of weekly docetaxel plus trastuzumab versus weekly paclitaxel plus trastuzumab in patients with previously untreated advanced nonsmall cell lung carcinoma. Cancer 104: 2149-2155, 2005

[36] Thall PF and Wathen JK: Bayesian designs to account for patient heterogeneity in phase II clinical trials. Curr Opin Oncol 20:407-411, 2008

[37] Rosner GL, Stadler W, and Ratain, MJ: Randomized discontinuation design: application to cytostatic antineoplastic agents. J Clin Oncol 20: 4478-4484, 2002

[38] El-Maraghi RH and Eisenhaurer EA. Review of phase II trial designs in studies of molecular targeted agents: Outcomes and predictors of success in phase III. J Clin Oncol 26: 1346-1364, 2008

[39] Ratain MJ, Humphrey RW, Gordon GB *et al.*: Recommended changes to the oncology clinical trial design: Revolution or evolution? Eur J Cancer 44: 8 – 11, 2008

[40] Booth CM, Calvert AH, Giaccone G *et al.*: Design and conduct of phase II studies of targeted anticancer agents: Recommendations from the task force on methodology for the development of innovative cancer therapies. Eur J Cancer 44: 25-29, 2008

[41] Centers for Disease Control and Prevention: Prevalence of stroke – United States, 2005. Morb Mort Wkly Rep 56: 469 – 474, 2007

[42] Alberts MJ: tPA in acute ischemic stroke. United States experience and issues for the future. Neurology 51: S53-S55, 1998

[43] Labiche LA and Grotta JC: Clinical trials for cytoprotection in stroke. NeuroRX 1: 46-70, 2004

[44] Gladstone DJ, Black SE, Hakim AM *et al.*: Toward wisdom from failure: Lessons from neuroprotective stroke trials and new therapeutic directions. Stroke 33: 2123-2136, 2002

[45] Stroke Therapy Industry Round Table: Recommendations for standards regarding preclinical neuroprotective and restorative drug development. Stroke 30: 2752-2758, 1999

[46] Philip M, Benatar M, Fisher M *et al.*: Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials. Stroke 40: 577 -581, 2009

[47] McLeod MR, van der Worp HB, Sena ES *et al.*: Evidence for the efficacy fo NXY-059 in experimental focal cerebral ischemia is confounded by study quality. Stroke 39: 2824-2829, 2008

[48] Stroke Therapy Academic Round Table II. Recommendation for clinical trial evaluation of acute stroke therapies. Stroke 32: 1598-1606, 2001

[49] Fisher M : Recommendations for advancing development of acute stroke therapies: Stroke therapy academic industry roundtable 3. Stroke 34: 1539-1546; 2003

[50] Fisher M and for the Stroke Therapy Academic Industry Roundtable IV: Enhancing the development and approval of acute stroke therapies: Stroke Therapy Academic Industry Roundtable. Stroke 36: 1808-1813, 2005

[51] Saver JL, Albers GW, Dunn B, *et al.*: Stroke Therapy Academic Industry Roundtable (STAIR): Recommendations for extended window acute stroke therapy trials. Stroke 40: 2594-2600, 2009

[52] Lees KR, Zivin JA, Ashwood T *et al.*: NXY-059 for acute ischemic stroke. N Engl J Med 354: 588-600, 2006

[53] Shuaib A, Lees KR, Lyden P *et al.*: NXY-059 in the treatment of acute ischemic stroke. N Engl J Med 357: 562-571, 2007

[54] Koziol JA andFeng AC: On the analysis and interpretation of outcome measures in stroke clinical trials: Lessons from the SAINT I study of NXY-059 for acute ischemic stroke. Stroke 37: 2644-2647, 2006

[55] Marshal G, Beaudouin V, Rioux P *et al.*: Prolonged persistence of substantial volumes of potentially viable brain tissue after stroke. A correlative PET-CT study with voxel-based data analysis. Stroke 27: 599-606, 1996

56 Baron J: Mapping the ischemic penumbra with PET: Implications for acute stroke treatment. Cerebrovasc Dis 9: 193-201, 1999

57 Fisher M and Gingsberg M: Current concepts of the ischemic penumbra: an introduction. Stroke 35:2657-2658, 2004

58 Heiss WD, Thiel A, Grond M *et al.*: Which targets are relevant for the therapy of acute ischemic stroke? Stroke 30: 1486-1489, 1999

59 Hacke W, Furlan AJ, Al-Rawi, Y *et al.*: Intravenous desmoteplase in patients with acute ischaemic stroke selected by MRI perfusion-diffusion weighted imaging or perfusion CT (DIAS-2): a prospective, randomized, double-blinded study. Lancet Neurol 8: 126-128, 2009

60 Hacke W, Albers G, Al-Rawi Y et al: The Desmoteplase in Acute Ischemic Stroke Trial (DIAS): A Phase II MRI-Based 9-Hour Window Acute Stroke Thrombolysis Trial With Intravenous Desmoteplase. Stroke 36: 66-73, 2005

61 Furlan AJ, Eyding D, Albers GW et al: Dose Escalation of Desmoteplase for Acute Ischemic Stroke (DEDAS): evidence of safety and efficacy 3 to 9 hours after stroke onset. Stroke 37: 1227-1231, 2006

62 Albers GW, Thijs VN, Wechsler L *et al.*: Magnetic resonance imaging profiles predict clinical response to early reperfusion: The diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study. Ann Neurol 60: 508–517, 2006

63 Olivot J-M, Mlynash M, Thijs VN *et al.*: Geography, structure and evolution of diffusion and perfusion lesions in diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE). Stroke 40: (published online prior to print), 2009

64 The North American Glycine Antagonist in Neuroprotection (GAIN) Investigators: Phase II studies of the dlycine antagonist GV150526 in acute stroke: The North American experience. Stroke 31: 358-365, 2000

65 Abciximab Emergent Stroke Treatment Trial (AbESTT) Investigators: Emergency administration of abciximab for treatment of patients with acute ischemic stroke: Results of a Phase 2 trial. Stroke 36: 880-890, 2005

66 Adams HP, Effron MB, Torner J *et al.*: Emergency administration of abciximab for treatment of patients with acute ischemic stroke: Results of an international phase

III trial – Abciximab in emergency treatment of stroke trial (AbESTT II). Stroke 39: 87-99, 2008

67 Ho PW, Reutens DC, Phan TG *et al.*: Is white matter involved in patients entered into typical trials of neuroprotection? Stroke 36: 2742-2744, 2005

68 Jonas S, Aiyagari V, Vieira D *et al.*: The failure of neuronal protective agents versus the success of thrombolysis for the treatment of acute ischemic stroke. Ann N Y Acad Sci 939: 257-267, 2001

69 Dyker AG and Lees KR: Duration of neuroprotection treatment for ischemic stroke. Stroke 29: 535-542, 1998

70 Saunders DE, Howe FA van der Boogaart *et al.et al.*: Continuing ischemic damage after acute middle cerebral artery infarction in humans demonstrated by short-echo proton spectroscopy. Stroke 26:1007-1013, 1995

71 Goldstein LB: Pharmacology of recovery after stroke. Stroke 21: 139-142, 1990

72 Grotta J: Neuroprotection is unlikely to be effective in humans using current trial designs. Stroke 33: 306-307, 2001

73 Lees KR, Davalos A, Davis SM *et al.et al.*: Additional outcomes and subgroup analyses of NXY-059 for acute ischemic stroke. Stroke 37: 2970-2978, 2006

74 Palesch YY, Tilley BC, Sackett DL *et al.et al.*: Applying a phase II futility study design to therapeutic stroke trials. Stroke 36: 2410-2414, 2005

75 MRI Collaborative Group: Proof-of-principle phase II MRI Studies in stroke: Sample size estimates from dichotomous and continuous data. Stroke 37: 2521-1525, 2006

76 Mandava P and Kent TA: A method to determine stroke trial success using multidimensional pooled control functions. Stroke 40: 1803-1810, 2009

77 EMEA: Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. Doc. Ref. CPMP/EWP/561/98 Rev. 1, 16 November 2006

78 Miller AB, Hoogstraten B, Staquet M, and Winkler A: Reporting results of cancer treatment. Cancer 47: 207-214, 1981

79 Therasse P, Arbuck SG, Eisenhauer EA *et al.et al.*: New guideline to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 92"205-218, 2000

80 Eisenhauer EA, Therasse P, Bogaerts J *et al.*: New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247, 2009

81 Goffin J, Baral S, Tu D *et al.*: Objective response in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. Clin Cancer Res 11: 5928 – 5934, 2005

82 Elhilali MM, Gleave M, Fradel Y *et al.*: Placebo-associated remissions in a multicentre, randomized, double-blind trial of interferon γ-1b for the treatment of metastatic renal carcinoma. B J U Intl 86: 613-618, 2000

83 Lavin PT: An alternative model for the evaluation of antitumor activity. Cancer Clin Trials 4: 451-457, 1981

84 Karrison TG, Maitland ML, Stadler WM *et al.*: Design of phase II cancer trials using a continuous endpoint of change in tumor size- Application to a study of sonaferib and erlotinib in non-small cell lung cancer. J Natl Cancer Inst 99: 1455-1461, 2007

85 Francart J, Lagrand C, Sylvester R *et al.*: Progression-free survival as primary endpoint for Phase 2 cancer trials. Application to mesothelioma-The EORTC Lung Cancer Group. J Clin Oncol 24:3007-3012, 2007

86 Freidlin B, Korn EL, Hunsberger S *et al.*: Proposal for the use of progression-free survival in unblinded randomized studies. J Clin Oncol 25:2122-2126, 2007

87 Adjei AA, Christian M and Ivy P: Novel designs and endpoints for phase II clinical trials. Cin Cancer Res 15: 1866-1872, 2009

88 Dhani N, Tu D, Sargent D *et al.*: Alternate endpoints for screening phase II studies. Clin Cancer Res 15: 1873-1882, 2009

89 Van Glabbeke M, Verweij J, Judson I *et al.*: Progression-free rate as the principal endpoint for phase II trials in soft-tissue sarcomas. Eur J Cancer 38: 543-549, 2002

90 Shankar LK, van der Abbeele A, Yap J *et al.*: Considerations for the use of imaging tools for the phase II treatment trials in oncology. Clin Cancer Res 15: 1891-1897, 2009

91 Kellof GJ and Sigman CC: New science-based endpoints to accelerate oncology drug development. Eur J Cancer 41: 491-501, 2005

92 Dalton SW and Friend SH: Cancer biomarkers – an invitation to the table. Science 312: 1165-1168, 2006

93 Brott TG, Adams HP, Olinger CP *et al.*: Measurements of acute cerebral infarction: a clinical examination scale. Stroke 20: 864-870, 1989.

94 Mahoney FI and Barthel DW: Functional evaluation. The Barthel Index. Md State Med J 14: 61-65, 1965

95 Schiemanck SK, Post MWM, Kwakkel G *et al.*: Ischemic lesion volume correlates with longer term functional outcome and quality of life of middle cerebral artery stroke survivors. Restor Neurol Neurosci 23: 257-263, 2005

96 Granger CV, Hamilton DB and Gresham GE: The stroke rehabilitation outcome study – Part I: general description. Arch Phys Med Rehabil 69: 506 – 509, 1988

97 Dromerick AW, Edwards DF and Diringer MN: Sensitivity to changes in disability after stroke: a comparison of four scales useful in clinical trials. J Rehabil Res Dev 40: 1- 8, 2003

98 Van Swieten JC, Koudstaal PJ, Visser MC *et al.*: Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19: 604-607, 1988

99 Schiemack SK, Post MWM, Witcamp TD *et al.*: Relationship between ischemic lesion volumes and functional status in the 2nd week after middle cerebral artery stroke. Neurorehabil Neural Repair 19: 133 – 138.

100 Murray GD, Barer D, Choi S *et al.*: Design and analysis of phase III tr4ials with ordered outcome scales: the concept of the sliding dichotomy. J Neurotrauma 22: 511-517, 2005

101 Lai S-M and Duncan PW: Stroke recovery profile and the modified Rankin assessment. Neuroepidemiology 20: 26-30, 2001

102 Jennett B and Bond M: Assessment of outcome after severe brain damage: a practical scale. Lancet 1: 480-484, 1975

103 Kasner SE: Clinical interpretation and use of stroke scales. Lancet Neurol 5: 603-612, 2006

104 Young FB, Weir CJ, Lees KR *et al.* Comparison of the National Institutes of Health Stroke Scale with disability outcome measures in acute stroke trials. Stroke 2005: 2187-2192, 2005

105 Sulter G, Steen C and de Keyser J: Use of the Barthel Index and Modified Rankin Scale in Acute Stroke Trials. Stroke 30:1538-154, 1999

106 Hacke W, Kaste M, Fieschi C *et al.*:. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). Second European-Australasian Acute Stroke Study Investigators. Lancet. 352: 1245–1251, 1998

107 Shinar D, Gross CR, Bronstein KS *et al.*: Reliability of the activities of daily living scale and its use in telephone interview. Arch Phys Med Rehab 68: 723-728, 1987

108 Goldstein LB and Samsa GP: Reliability of the National Institutes of Health stroke scale: extension to non-neurologists in the context of a clinical trial. Stroke 28: 301-310, 1997.

109 Lyden P, Raman R, Liu L *et al.*: NIHSS training and certification using a new digital video disk is reliable. Stroke 36: 2446-2449, 2005

110 Josephson SA, Hills NK and Johnston SC: NIH stroke scale reliability in rating from a large number of clinicians. Cerebrovasc Dis 22: 389-395, 2006

111 Wilson JTL, Hareendran A, Grant M: Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin scale. Stroke 33:2243-2246, 2002

112 Newcommon NJ, Green TL, Haley E *et al.*: Improving the assessment of outcomes in stroke: Use of a structured interview to assign grades on the modified Rankin Scale. Stroke 32: 2021-2028, 2003

113 Wilson JTL, Hareendran A, Hendry A *et al.*: Reliability of the modified Rankin scale across multiple raters: Benefits of a structured interview. Stroke 36:777-781, 2005

114 Shinehana Y, Minematsu K, Armano T *et al.* Modified Rankin scale with expanded guidance scheme and interview questionnaire: Interrater agreement and reproducibility of assessment. Cerebrovasc Dis 21:271-278, 2006

115 The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group: Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 333: 1581-1587, 1995

116 Tilley BC, Marler MD, Geller NL *et al.*: Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and stroke t-PA trial. Stroke 27: 2136-2142, 1996

117 Saver JL, Johnston KC, Homer D *et al.*: Infarct volume as a surrogate or auxiliary outcome measure in ischemic stroke clinical trials. Stroke 30: 293-298, 1999

118 The RANTTAS Investigators: A Randomized Trial of Tirilazad Mesylate in Patients With Acute Stroke (RANTTAS). Stroke 27:1453-1458, 1996

119 Luby M, Bykowski JL, Schellinger PD *et al.*: Intra- and interrater reliability of ischemic lesion volume measurements on diffusion-weighted, mean transit time and fluid attenuated inversion recovery MRI. Stroke 37: 2951-2956, 2006

120 Savitz SI, Fisher M: Future of neuroprotection for acute stroke: In the aftermath of the SAINT trials. Ann Neurol 61: 396-402, 2007

121 Van der Worp HB, Kapelle LJ, Algra A *et al.*: The effect of tirilizad mesylate on infarct volume of patients with acute ischemic stroke. Neurology 58: 133-135, 2002

122 Tirilazad International Steering Committee: Tirilazad mesylate in acute ischemic stroke: a systematic review. Stroke 31: 2257-2265, 2000

123 Warach S, Kaufman D, Chiu D *et al.*: Effect of glycine antagonist gavestinel on cerebral infarcts in acute stroke patients, a randomized, placebo-control trial: the GAIN MRI substudy. Cerebrovasc Dis 21: 106-111, 2006

124 Kidwell CS, Lees KR, Muir KW *et al.*: Results of the MRI substudy of the intravenous magnesium efficacy in stroke trial. Stroke 40: 1704-1709, 2009

125 Davis SM, Donnan GA, Parsons MW et al: Effects of alteplase beyond 3 h after stroke in the Echoplanar Imaging Thrombolytic Evaluation Trial (EPITHET): a placebo-controlled randomised trial. Lancet Neurol 7: 299-309, 2008

126 Menezes NM, Ay H, Zhu MW *et al.*: The real estate factor: Quantifying the impact of infarct location on stroke severity. Stroke 38: 194-197, 2007

127 EMEA: Points to consider on clinical investigations of medicinal products for the treatment of acute stroke. CPMP/EWP/560/98, 20 September 2001

128 Wintermark M, Albers GW, Alexandrow JR *et al.*: Acute stroke imaging roadmap. Stroke 39: 1621-1628, 2008

129 Wardlaw JM: Surrogate outcomes: a cautionary note. Stroke 40: 1029-1031, 2009

130 Gwyther SJ, Aapro MS, Hatty SR *et al.*: Results of an independent oncology review board of pivotal trials of gemcitabine in non-small cell lung cancer. Anticancer Drugs 10: 693-698, 1999

131 Ford R, Schwartz L, Dancey J *et al.*: Lessons learned from independent central review. Eur J Cancer 45: 268-274, 2009