Published in final edited form as:

Curr Alzheimer Res. 2008 August ; 5(4): 346–357.

Alzheimer's Disease Drug Development in 2008 and Beyond: Problems and Opportunities

Robert E. Becker^{*} and **Nigel H. Greig**

Drug Design & Development Section, Laboratory of Neurosciences, Intramural Research Program, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA

Abstract

Recently, a number of Alzheimer's disease (AD) multi-center clinical trials (CT) have failed to provide statistically significant evidence of drug efficacy. To test for possible design or execution flaws we analyzed in detail CTs for two failed drugs that were strongly supported by preclinical evidence and by proven CT AD efficacy for other drugs in their class. Studies of the failed commercial trials suggest that methodological flaws may contribute to the failures and that these flaws lurk within current drug development practices ready to impact other AD drug development [1]. To identify and counter risks we considered the relevance to AD drug development of the following factors: (1) effective dosing of the drug product, (2) reliable evaluations of research subjects, (3) effective implementation of quality controls over data at research sites, (4) resources for practitioners to effectively use CT results in patient care, (5) effective disease modeling, (6) effective research designs.

New drugs currently under development for AD address a variety of specific mechanistic targets. Mechanistic targets provide AD drug development opportunities to escape from many of the factors that currently undermine AD clinical pharmacology, especially the problems of inaccuracy and imprecision associated with using rated outcomes. In this paper we conclude that many of the current problems encountered in AD drug development can be avoided by changing practices. Current problems with human errors in clinical trials make it difficult to differentiate drugs that fail to evidence efficacy from apparent failures due to Type II errors. This uncertainty and the lack of publication of negative data impede researchers' abilities to improve methodologies in clinical pharmacology and to develop a sound body of knowledge about drug actions. We consider the identification of molecular targets as offering further opportunities for overcoming current failures in drug development.

INTRODUCTION

Drugs can fail in CTs but, by not showing proper diligence in the design, monitoring, analysis, and interpretation of a CT, investigators can fail drugs. In recent reports, investigators identify how systematic and random measurement errors can potentially undermine the demonstration of efficacy in Alzheimer's disease (AD) and depression clinical trials (CT) [1–4]. The dependence on rating scales as outcome measures seems the problematic feature common to clinical research in these fields. With the assistance of Elemer Piros (Senior Biotechnology Analyst, Rodman and Renshaw) and Neil Buckholz (Chief, Dementias of Aging Branch, National Institute on Aging, NIH) we identified nearly 100 compounds with over 40 different mechanisms of action that had been considered as potential treatments for AD. We found 20

^{*}Address correspondence to the author at the Drug Design & Development Section, Laboratory of Neurosciences, 5th floor BRC Building, 251 Bayview Blvd., Baltimore, MD 21224, USA; E-mail: rebecker2008@comcast.net.

compounds with preliminary evidence for benefits in AD preclinical studies and in Phase II but that failed to show consistent successes in Phase III CTs. We could not obtain exact information about many compounds due to commercial sponsors refusing to publish or release to reviewers information about negative outcomes from their drug development. Our initial estimate is that this affects some 40% to 50% of the compounds and is a problem that, likewise, is encountered by others [5]. Our tabulated results of this review of AD compounds may thus never be entirely complete, and will form the basis of a separate article.

Using available reports about error effects in CTs, failed drug developments in AD and depression, and our own experiences of developing drugs for AD, we identified six issues that affect the success, costs, and subsequent clinical applications of CTs [6–11]. In this article, we discuss possible steps to overcome these impediments to successful AD drug development. Specifically, we address how to avoid methodologically flawed trials and consequent failures in development of potentially viable AD drug candidates. We consider the preclinical grounding needed to assure effective dosing of AD drugs; overcoming problems with the reliability of rated evaluations of CT research subjects; the need for close monitoring to assure quality of data obtained at research sites; how CTs can provide resources for effective clinical applications of results by practitioners; the importance of effective disease modeling in the CT; and the implications for effective research design for AD CTs.

AVOIDING METHODOLOGICALLY FLAWED TRIALS AND FAILURES OF PHARMACOLOGICALLY VIABLE AD DRUGS DURING DEVELOPMENT

Many different cholinesterase inhibitors (ChEIs) have shown both preclinical properties that predict efficacy in AD CTs and then in clinical trials (CTs) demonstrated efficacy with statistically significant improvements of drug treated patients compared to placebo treated patients [12,13]. Hence, the inhibition of acetylcholinesterase can now be considered a valid drug target consequent to the approval of four agents of this class by the U.S. Food and Drug Administration (US FDA) (tacrine (*Cognex*) 1993, donepezil (*Aricept*) 1997, rivastigmine (*Exelon*) 2000) and galantamine (*Reminyl*) 2001). Recently, we were given the opportunity to review two ChEIs with preclinical profiles consistent with these approved drugs; yet, these drugs failed in large CTs and were abandoned or seem on the verge of being abandoned by developers. In addition to reviews under confidentiality agreements of these two compounds, we also consider our direct preclinical and clinical research experience with heptylphysostigmine, phenserine, and metrifonate [14–19]. Heptylphysostigmine failed in CTs due to toxicity. Do the failures of phenserine and metrifonate reveal flaws in the abilities of pre-clinical studies to characterize cholinesterase inhibitors as clinical candidates or did drug development planning and the environment for conducting AD CTs fail to provide an unbiased test of these AD drugs? If current AD CT conditions can contribute to failed drug development, then new classes of drugs for AD may also be at risk [20]. The problem faced by developers of AD drugs is that the already failed drugs may, indeed, be drug failures; however, conditions of development may also have caused these failures and could haunt new drug developments in AD. As we have already published and from publications addressing failures in drugs for depression, we find adequate evidence to suggest that current CT practices may cause or significantly contribute to commercial failures of AD drugs. We review this evidence to remove methodological impediments that may deny AD drugs fair and unbiased CT tests of efficacy.

Becker *et al.* [21–23] and Perkins *et al.* [24] found a mean of multiple ratings necessary to overcome unreliability that interferes with the power in CTs. Lachin [25] calls for all studies to report the reliability of measurements as a step towards improved CT designs. Becker [1] presents evidence for how unexpectedly high levels of variance in data sets can explain the failures of CTs to support efficacy, and reviews work by statisticians and investigators of depression CTs who report how conditions of drug testing in CTs can account for failures to

demonstrate statistically supported efficacy [26–29]. Thase [30] calls for methods to improve the ability to discriminate between active antidepressants and placebo as needed to avoid dependence on meta-analyses for trials using outcome raters. Muller and Szegedi [31] emphasize that lack of power due to unreliable assessments risks false-negative reports and ethical consequences. CTs dependent on rated outcomes are widely suspect as not providing fair efficacy trials of drugs.

Demitrack *et al.* [32] found that, for some raters, unreliability of ratings in a depression trial could not be improved with extended training and investigators remain undecided about how problems of rater unreliability in CTs can be overcome [26,33,34]. It is evident from these sources that no simple formula assures the required careful preparations, design, training, implementation, analyses, and interpretations required for success in a CT. Becker's [1] review reveals the vulnerabilities of CTs to inaccuracy and imprecision in data, and both justifies and describes steps to be taken in order not to repeat these sources for failure in future CTs. In the CTs of ChEIs used as examples in Becker [1], limitations inherent in the drugs appear not to account for the failures of CTs to confirm efficacy. Each of the ChEIs considered in Becker [1] were shown in preclinical research to be available in sufficient concentrations in brain and came from a class already shown to produce the molecular-kinetic changes in acetylcholine metabolism needed to improve cognitive performance and behavior in AD [9,22,23,35,36]. The failed CTs did not report higher rates of interfering adverse events than occurred in the successful CTs. Each of the ChEIs showed in preclinical studies the required pharmacokinetic and pharmacodynamic properties. For these drugs the environment of CTs seemed to strongly affect outcomes. An initial large commercial CT of phenserine failed due to excessive variance in observations [37] while a smaller academic and a more closely monitored, commercially sponsored CT yielded less variance and statistically significant efficacy (36, 38). From experiences in other fields and developing evidence in AD drug testing, one is drawn to the conclusion that investigators, by choices in how they would develop their drugs, can fail their drugs by not providing fair tests for efficacy and adequate safeguards against compromises to safety.

This view, that in drug development and in CTs investigators can fail to provide conditions that fairly test their drugs, encourages reconsideration of preclinical preparations and aims of initial clinical phases of drug development. For example, recent CTs of immunization against amyloid were interrupted by adverse events that could be interpreted as cautions against investigators proceeding too quickly to large scale human studies [39]. *In vitro* and animal models for the mechanism of action of the drug need to anticipate how investigators can avoid possible adverse events that in later stages of clinical development can interfere with efficacy or threaten humans sufficiently to require termination of the human research or withdrawal of a New Drug Application (NDA) approval by the FDA. Drug development can usefully be thought of as a process of learning how to use a drug therapeutically as well as evidencing efficacy and safety in use. Drugs may be failing in both CTs and after NDA approvals from inadequately disciplined development. To overcome these lapses investigators, during preclinical *in vitro* and animal modeling of the drug's activities, can attempt to identify and avoid threats to success in CTs. We suggest this view to replace more passive or less intense inquiries during CT preparations. An aggressive yet cautious and conservative drug development seeks 'no surprises' in human phases; yet, achieves this goal while at the same time facilitating rapid and effective preparations for and then completion of clinical Phases I, II, and III CTs and uneventful Phase IV and prescription uses.

Often these aims of anticipating difficulties in CTs are not within our grasp as AD investigators; however, the costs from dosing failures, adverse events, and other factors interfering during CTs argue for efforts to exclude the unexpected prior to entering CTs. As we argue below, if investigators do not more actively seek solutions to limitations in drug development planning

and CTs we will not find solutions to the current limitations that negatively impact the effectiveness and costs of AD CTs [1]. Without a ‘no surprises’ quality control over prospective CT conditions, history teaches us to anticipate often unpleasant and costly surprises.

EFFECTIVE DOSING OF THE DRUG PRODUCT

Two examples illustrate how insufficient controls over dosing of a drug can undermine CT success. In a recent, 48-month, multi-center, double-blind, placebo-controlled CT studying 1018 persons with Mild Cognitive Impairment (MCI) at risk for AD [40], in spite of expert inputs into the design and execution of the research, investigators allowed sub-therapeutic doses of active drug to be tested against placebo [41]. At the other end of the spectrum of dosing, a commercially sponsored series of CTs testing the cholinesterase inhibitor, metrifonate, resulted in sufficiently serious adverse events to cause the drug development to be abandoned after completion of Phase III with demonstration of efficacy in two multi-center CTs, and submission of the NDA [16]. In this latter situation, there seems reason to associate the appearance of adverse events with the commercial sponsor’s choice to dose this ‘irreversible’ enzyme inhibitor on a daily basis and to do so without independently established restraints against dosage increases [42–44]. Even though the cholinesterase inhibition associated with the drug and its active metabolite, 2,2-dimethyl dichlorovinyl phosphate, is stable and may persist for in excess of a month [45], the ChEI used in these studies elicits only benign short-term adverse events that decrease if one allows time for the patient to accommodate to each dosing increase [8,46]. This accommodation by patients to adverse drug effects allowed commercial sponsors of development to use increased doses of the drug in their later studies [42]. In our research, this drug toxicity was shown to be associated with the highest levels of dosing reported in these studies [47]. Unfortunately, because of the proprietary ownership of human research results allowed under business law, full public access to these data and an unknown number of other negative CT results are not available for review [48].

In contrast to the commercially sponsored CTs, in an academic setting, in earlier development of metrifonate for use in AD carried out prior to its licensing for commercial development, investigators completed two Phase III clinical trials using weekly dosing designed, following *in vitro* studies, animal modeling and a Phase I–II human study, to avoid effects from excessive accumulated enzyme inhibition and to comply with an optimal dose-response relationship identified in the Phase I–II trial [21,22,49]. In drug exposure to metrifonate of approximately 150 patients for up to 5 years, these academic investigators encountered no serious adverse events even though one patient, unsupervised by her caretaker for three days, ingested a three months supply of her drug.

These dosing anecdotes provide figurative bookends for the concept of a range of acceptable and optimal dosing for a drug. We interpret the above and similar problems with containing dosing within a safe and effective range as illustrating the importance for investigators that they obtain independent evidence to establish and defend the boundaries for optimal dosing of their investigational drug. Under conditions of Phases III and IV, inadequate or toxic levels of dosing can occur because investigators unwittingly compromise existing dosing guidelines to achieve the aims of their Phase III or IV study. In the MCI example it appears that an emphasis on obtaining data from an adequate patient number (N) within the time allowed to complete the study may have taken priority over and compromised dosage guidelines. In the metrifonate example it appears, from discussions in symposia, that cautions about the commercial developer’s already aggressive dosing were overlooked as hopes prevailed in development that higher doses would improve the magnitude or duration of drug effects and provide an advantage over competing ChEIs. Important considerations thus arise for commercial developers: does positioning a new drug in relation to those already available for prescription undermine the efficacy and safety available from the new compound? One does not develop a drug into an

agent with a level of efficacy that is better tolerated than competitor drugs or an agent with superior efficacy that is similarly tolerated; a drug reveals itself in the developmental studies and desirable but not inherent features are sometimes elicited by carefully defining conditions of its use. The regulatory required studies are not hurdles to be surmounted but one context within which investigators explore drug properties.

Sound Phase II studies focused on finalizing dosing parameters are often put aside today to design Phase II–III combined studies. Phase II studies need designs that focus on confirming dosing parameters without competition from other aims. The properties of the drug in patients, not desires of or development related pressures on sponsors and investigators, need to control dosing. In both Phases I and II, efficacy data can be usefully part of the studies which focus on the interrelationships of dosing, efficacy, and safety but, unlike Phase III, not a primary aim in these Phases. Phase III confirms the efficacy of the optimal dosing range established in Phases I and II. As the two examples offered illustrate, priorities to achieve Phase III or IV CT aims can easily undermine the safety or efficacy of the drug if optimal dosing requirements are not empirically established, set as study conditions prior to the start of Phase III, and observed in all CTs.

Other factors shown to influence success and failure in CTs, as demonstrated in the MCI study, need also to be as fully as possible understood for their interference prior to initiating a CT [41]. One among many examples of confounding problems noted as interfering in the MCI trial [41] is the high rates of other diseases in the prospective population, the consequent use by the study sample of prescribed medications with associated adverse events, and the resulting potential for high rates of adverse events, over 90% in the MCI study in both treatment and placebo arms. The statistical assumptions behind CT designs open the investigational drug to association with high rates of adverse events and to dosage reductions by investigators blind to whether the adverse events show patterns characteristic of an association with the research drug or with the population being studied. An increased integration of clinical phases, as we recommend for consideration below, provides sponsors opportunities to evaluate during earlier phases any potentially interfering effects from available study populations and to design Phase III with accommodations and protections against subject characteristics and conditions at research sites interfering with fair CT testing of the investigational drug. In practical terms, smaller Ns seem needed in AD CTs to allow investigators selecting subjects to exclude those using other medications with adverse effects that could interfere with the study meeting its required dosing conditions.

For many investigational drugs being tested for efficacy in AD, there are currently no immediate or short-term indicators or markers that investigators can consult to estimate the long-term efficacy of the drug given at the specific dosing being studied. We find having available one or more well-characterized animal models of dose to targeted molecular response relationships to be essential to our understanding the dose-response implications of Phase I and II subject responses [6,10]. For example, in preparation for administration of metrifonate to human subjects in a Phase I–II study with aims to establish an optimal dose, to characterize the dose-response curve, and to demonstrate the safety from toxicity at any dosing level or patient response disposition that could be encountered clinically from accidental overdosing, from deviant pharmacokinetics, from low body weight, and so forth, we studied indicators of pharmacokinetics and pharmacodynamics in over 5 species and using over 5 different routes of administration [6]. In addition we studied the *in vitro* kinetics of enzyme inhibition using animal and human cholinesterases [8,18]. These experiences established patterns of relationships among cholinesterase inhibition levels in red blood cells and plasma, in cerebrospinal fluid, physiological markers, adverse events, brain cholinesterase inhibition, and so forth. We could then use changes in parameters that we could study in humans to estimate levels of brain cholinesterase inhibition in our human subjects. Informed by these estimates

we could aim our proposed Phase I–II dosing at achieving specific ranges of change in our molecular targets brain acetylcholinesterases and butyrylcholinesterases and evaluate our data in terms of estimated effects at the drug target, an effect not available to direct measurement. The clinical responses we could measure in patients short-term became indicators of change at molecular targets

Prior to entry into the human species, we view as essential that investigators gain knowledge of animal models adequate to estimate targeted responses in humans with adequate reliability and certainty. This use of modeling supports investigators to develop independent criteria for human dosing prior to Phase III and to explore for possible toxicity at dosing levels that might clinically occur above the optimal dose. One is most fortunate if sufficiently direct and quantitative indicators are available for the molecular target, such that the target itself can be the dependent variable rather than clinical behaviors. Given the molecular mediation of drug effects, the molecular specificity of drugs currently under development in AD, the limitations on the reliability of clinical ratings as we discuss below, and the rapidly evolving development of new distinctive and reliable biomarkers for AD pathologies, we conclude that drug development evaluated using drug activities against molecular disease mechanisms may offer important advantages over current dependence on rated clinical responses of research subjects.

RELIABLE EVALUATIONS OF RESEARCH SUBJECTS

Becker [1] compared four clinical trials and found that variance could account for the successful demonstrations of efficacy in two trials and the failure of two other trials. Each of the failed trials used larger Ns than did the two successful CTs. Larger Ns increase the statistical power of the trial to detect group differences. Unfortunately, in spite of their using larger Ns, higher levels of variance appear to account for the failure of two trials to reach statistical significance.

Becker [1] focused mainly on one source for error variance, excessive random measurement error, and on methods to achieve sufficient control over this source of variance, such that statistical significance could be shown for treatments with relatively small numbers of patient subjects. The two Becker *et al.* [21,22] CTs used to demonstrate the practicality of small Ns in CTs [1], compared to other published AD CTs, attained relatively tight standard deviations for patient ratings which standard deviations were further reduced by averaging three observations to obtain each data point entered into the data analysis [23]. As reported, but not discussed in detail in Becker [1], recently a series of articles in multi-center studies of depression, documented the presence of highly consequential inaccuracies in evaluations of research subjects completed by over 50% of trained research raters [2–4,50–52]. As a result of the prevalent inaccuracies and imprecision in ratings in which human judgments generate the data obtained from patient observations, it is our view that sponsors of CTs that use rating scales for AD, depression, neurological conditions, and so forth, must assure that systematic errors, bias, carelessness during periods of observation, inattention to the study protocol and imprecision in ratings do not interfere with or undermine findings of statistical significant differences for the investigational treatment. Unfortunately, the required reliabilities may not be achievable using either trained research raters or practitioners working with individual patients.

In Becker [1] the partitioning of sources of variance and estimations of interventions aimed at controlling random measurement errors do not take into account problems of rater inaccuracy or the advantages gained with evaluations of molecular target changes (e.g. biomarkers) as the dependent variable. A design with molecular target changes as the dependent variable should have inherently greater power given that the variance associated with the molecular target response can have tighter error distributions given laboratory and not human measurements, and will be independent of the variance added if clinical responses are rated or estimated by

practitioners. Given the problems with errors in rating scales, validated laboratory-based outcome measures should contribute to reductions in sizes of CTs and the successful completion of CTs and management of drugs by practitioners providing patient care [29]. Thus as the partitioning of variance model [1] suggests, measurements of a molecular target as the dependent variable should increase power over clinical measure outcomes both by the absence of the latter clinical measures' variance components and by the probably smaller standard deviation to change score proportions for measurements of a physical target as the dependent variable.

The clinical importance of molecular changes clearly remains a requirement for an NDA. Engelhardt *et al.* [2] describe practical foundational environmental conditions needed for raters to operate within if investigators are to prevent inaccuracies by raters from adversely interfering with CT analyses. Unfortunately, biases, inaccuracies, and other systematic errors by raters differ from random measurement errors such that accuracy requires external controls—external 'Gold Standard' assessments—to detect and correct rater inaccuracies [1,9]. Excessively random measurements can be detected by analyses internal to the CT data set, or by using data obtained independent of the CT [23]. Systematic errors, for example the inability of up to 50% of raters in two multi-center depression CTs to detect drug effects in patients, can in some cases be suspected from patterns in data but are established as present only by comparison of suspected raters' ratings with ratings by expert (Gold Standard) raters [2,34, 53].

In the model design for a drug development presented below, we recommend practical and ongoing training, evaluation, testing, and selection of raters during Phases I and II. These revised procedures for preparing for CTs can use indicators of random measurement errors and systematic errors and comparisons with Gold Standard Raters, to assess and insure the quality of data in Phases I and II. By screening out incompetent raters, sponsors of studies can create conditions favorable for the success of Phase III studies. These steps, while unusual, seem both economical and scientifically indicated for consideration to avoid the unnecessary failures of Phase III CT studies and the abandonment of otherwise effective drugs, the ambiguity this introduces into our understanding of the role of molecular mechanisms in AD, the economic losses incurred by Phase III failures, loss of time under patent protection due to recruiting unnecessarily large Ns to counter uncontrolled variance, the business consequences of failures and fears by potential investors of possibilities of repeated failures, and so forth.

Of course, no training or supervision of raters can overcome inherent limitations in measurement. For example, there is considerable reason to believe that, for conditions such as Parkinson's [54] and Alzheimer's diseases [55], relevant neurological circuitry that will be lost may be lost before definitive clinical signs and symptoms of disease appear. This could mean that ceiling effects of AD cognitive and behavioral tests may reflect only the accelerating onset of clinical illness and the later floor effects reflect the absence of adequate numbers of still functional neurons to support cognition and behaviors. In effect, some considerable percentage of function of relevant neurons may be lost while tests can not distinguish cognitive and behavioral differences between pre-AD and normals, consequent to the remarkable redundancy of the brain. In this scenario, most clinically detectable abnormalities occur during the final stages of neuronal loss. Preclinical functional and anatomical neuronal losses may explain problems using clinical rating scales as outcome measures for studies of MCI or other early AD related states. Plagued already by problems of reliability, rated changes in clinical status during pre-Alzheimer's MCI may occur only after proportionally greater losses of neurons compared to post-AD diagnosis. As an example, Clark *et al.* [56] found ceiling effects interfered with assessing the clinical status of mild AD patients and Feldman *et al.* [57] found floor effects with scales not especially designed for severe AD patients.

Indeed, floor and ceiling effects in commonly used AD rating scales may be not artifacts of scale construction, but reflections of non-linear brain-behavioral responses to functional and actual neuronal losses. We hence have good reasons to worry that clinical signs and symptoms allow study of only a small segment of the pathological processes underlying disorders such as AD.

Moreover, how clinical sign and symptom constructs for MCI and AD translate into multiple cultures and languages remains yet another relevant unknown [41]. Large Phase III trials are often cross-cultural, pooling patients from often widely divergent backgrounds and prior access to medical care.

EFFECTIVE IMPLEMENTATION OF QUALITY CONTROLS OVER DATA AT RESEARCH SITES

Typically, in a commercially sponsored drug development, preclinical studies are carried out in commercial laboratories isolated from communications with clinical investigators, Phase I testing is carried out at one set of clinical research sites, Phase II at a second set of sites, and Phase III at a third set of sites. The sponsor does not develop an ongoing relationship to the research sites on which the sponsor depends for crucial data and, because the conceptual focus in each step is on obtaining the data required for an investigation new drug (IND) application and then NDA, the sponsor typically receives a ‘bottom line’ report containing regulatory required data with little information about how this drug affects the diversity of patients or subjects exposed, except for case reports for serious adverse events.

We ask “Do current practices for carrying out preclinical, Phases I, II, and III studies make the most effective use of resources potentially available to support the demonstration of efficacy for a new drug?” Our answer is an emphatic “NO!” More constructive uses can be made of basic scientists to support clinical drug development and of Phases I and II to meet both current regulatory and pharmacological needs and to prepare for success in Phase III. One important step is to facilitate interactions between laboratory and clinic to address issues surfaced by studies in the human species. A second priority is to use preclinical and early clinical phases to insure the reliability of clinical data.

If the implications of inevitable problems with reliability of clinical data raised by Becker [1], Engelhardt *et al.* [2], Kobek [3], and Cogger [4] and others are correct, then it would be in a sponsor’s interests to use all available opportunities to assure the quality of data obtained by their clinical raters. Phases I and II provide just that opportunity to a sponsor at very little additional cost. Phases I and II can be used to train and screen investigators who will later become involved in Phase III (Figs. 1 and 2). If the problems of establishing optimal dose and the advantages of doing so are dependent on the availability of bench and animal models of drug kinetics and dynamics, as we have proposed above, then it is also in the sponsor’s interests to have basic scientists, experienced with the drug from pre-clinical studies, available to interact with, consult with, and even carry out further *in vitro* or animal assessments for clinical investigators to address issues of drug effects in the human species that might unfold during clinical development. Each of these two ongoing supportive environments—increased preclinical-clinical phase integration and site continuity during Phases I, II and III to improve rater performance in CTs—can be available if sponsors plan for their drug’s development by not using a linear, stepwise completed plan of current development. More useful is an overall conception of drug study and development, as carried out in various species with the aim of completing required regulatory steps but also, as developing investigators progressively more experienced with the drug and more skilled in the data acquisition that will be crucial to the success of Phase III.

We therefore recommend that sponsors consider institutionalizing in their drug development processes first, relationships between basic scientists who will develop the animal models for drug dosing in human species and those who will use these models and, second, a continuity of sites and personnel who will carry out Phases I, II, and III. A model proposal for this continuity with sites is presented in Figs. (1 and 2) and is discussed below.

PROVIDING THE RESOURCES FOR EFFECTIVE CLINICAL USE BY PRACTITIONERS

The most efficacious or well marketed drug will not serve its commercial sponsor or the public well if the drug is not used at effective doses or if patients do not continue to use the drug because of interfering adverse events or unrelated experiences perceived as interfering adverse events (see for example Feldman *et al.* [40] and Petersen [41]). Becker [1,9,58–60], in a series of articles, outlined methods to assess the effectiveness of specific doses of drugs, effectiveness in individuals, and helpful to assessing whether adverse events in an individual are drug related. These interventions are not currently practical using rating scales in clinical practice given the evidenced inadequacy of clinical assessments in most patient care dependent on the clinician's evaluation of the patient's reports [2]. If well-trained, paid raters frequently perform inadequately, or do not have the tools to control random measurement error interference, then it is reasonable to expect even greater interfering errors with 'eyeball' clinicians' judgments not grounded with rating scales and protocols for their administration. However, this does not mean that clinical decision rules based on laboratory clinical assessments could not contribute both to establishing a unique position in the market for a new drug, to countering hasty decisions by providers or patients to abandon treatment because of not obvious efficacy or apparent adverse events, and, most importantly, to the quality of care provided by medications [1,9]. These latter seem especially important in many chronic diseases where drugs provide no immediately obvious benefits for patients, but only abstract promises of long-term benefits and often uncomfortable adverse effects.

We therefore suggest, especially for sponsors of drugs with molecular targets, that drugs be developed to provide target-controlled clinical decision rules governing use [9]. This reassures the provider and consumer with empirical evidence of efficacy and lessens abuses from dosing that may lead to adverse events or unwarranted changes of drug. Adverse events may cause the patient to refuse medication, the physician to change the prescription or, if prevalent and serious, the FDA to review the NDA. Unfortunately a number of drugs recently lost FDA approval because of adverse events during prescription use. Clinical pharmacology better serves drug development if it causes investigators to recognize how drugs are inherently potentially dangerous and that mastery of the conditions of use, not inherent properties of the drug, assure safety and efficacy. We mistakenly see therapy as an inherent property of a drug when it is only a potentiality realized by carefully controlled and monitored use of the drug. The wider population exposed following a drug's approval brings on a further set of potential problems [61].

Sponsors can use post-double blind and other open drug treatment data to develop evidence for long-term effectiveness. With 'placebo periods' to demonstrate immediate drug efficacy, test data of effectiveness against molecular targets, and so forth, 'controlled' assessments of long-term outcomes can become available from open phase data and placebo arms can be avoided in some CTs [1,9]. A CT tested practice model for evaluating drug effectiveness in an individual using a period or periods of placebo administration or drug withdrawal can arm the practitioner with evidence of effectiveness needed to adjust dosing and to support the patient's continued use of the treatment. In this regard, a design for placebo period evaluations during open treatment in research or clinical practice can be found in Becker [1]. Due to regulatory priorities, CTs are regarded as required tests of efficacy and sponsors neglect the equally

important issues of demonstrating effectiveness in individual patients and providing models for rational clinical management of treatment in patients [1].

EFFECTIVE DISEASE MODELING

In two recent AD double-blind clinical trials placebo groups showed persistent secular trends of improvement during the double-blind phase [37; unpublished data]. In each of these trials the group improvements had a less than 5% chance occurrence based on the standard deviation for placebo group responses in CTs carried out in the 1980s and 90s and historical studies of AD. Any trial that does not replicate the established clinical disease lacks credibility as a test for treatments aimed at the disease. Based on the known pathology of AD, the specificity of diagnostic criteria, the past clinical courses of untreated AD, of placebo-treated AD, and of long-term drug treated AD patients, only random variation and bias can account for placebo group improvements. Given the evidence of inaccuracy and bias surfaced by Woods *et al.* [27], Engelhardt *et al.* [2], Kobek *et al.* [3], and others in depression studies, placebo group improvements can be taken as evidence of error interference in the data obtained in a study until proven otherwise. This evidence of bias and error undermines the entire data set. Consequently, a sponsor will want to avoid the presumed bias that provides placebo group improvements in AD CTs.

EFFECTIVE RESEARCH DESIGN

The aims of the considerations raised already—the relationship of CT success to dosing, measurement precision and accuracy, flexible uses of professional resources over the whole course of drug development, standardization and quality control at research sites, and consideration of prospective drug effectiveness for individuals aided by clinical decision rules—are to support sponsors of drug development to avoid problems that may account for drugs failing in development or being underused or withdrawn from prescription use because of inadequate guidance and controls over use by practitioners. As discussed, for drugs evaluated principally using the clinical judgments of practitioners and researchers, rather than with objectively quantifiable outcomes, much higher risks of error intrusions dispose development to failure. The next and following sections present model proposals to overcome these risks of failure. We present these models not as recommended for direct applications but as demonstrations that alternatives to current drug development practices are, indeed, available. In these models, drug development designs are not merely responsive to minimum regulatory requirements and, from fragmentation of development into steps, piecemeal in how they meet those requirements. These models and the discussions above aim to provide an overall concept of drug development as an optimal use of resources anticipating, and taking preventive action to avoid, difficulties in AD drug development. Experience seems to prove that risks to efficacy and safety demonstrations are almost inevitable unless sponsors and investigators remain vigilant to the possibility that, in a complex drug development in any way permissive of wrongs, if something can go wrong, something will go wrong.

A MODEL DRUG DEVELOPMENT CONCEPT

Aims

The following Model Drug Development Concept attends to four specialized areas to avoid failures that are present under current drug development practices. Implicit in these steps is the recognition that regulatory criteria for an IND and NDA and discussions of drug development and of CTs in the literature leave the impression that a sponsor plans for and overcomes a series of hurdles to qualify for an IND and NDA. In reality, success requires much reverse reasoning. For example, the efficacy and safety required in Phases III and IV require dosing controls that avoid sub-therapeutic or toxic doses in individuals. This requirement sets the Phase II goals of

determining a safe and effective dosing range. Phase II goals set Phase I priorities to achieve a safe dosing range. Phase I and II require *in vitro* studies and preclinical animal models since drug concentrations and activities at targets in humans can not be determined but must be assumed from indirect indicators. Problems with rater performance require preparations for Phase III to guard against failures and so forth. The four foci chosen to implement this 'reverse planning for no surprises' are:

1. To execute *in vitro* laboratory and animal model supports for dose finding activities in human Phases I and II.
2. To enable interactions of basic and clinical researchers to support human phases of drug development.
3. To design and execute clinical studies to reach maximum power without excessive Ns.
4. To use Phases I and II to test outcome measure administrations for reliability, to recruit sites, and to train, evaluate, and screen site evaluators for participation in Phase III.

This Model Drug Development Concept, while it specifically addresses the factors limiting the effectiveness of current CTs to meet research and regulatory aims, does appear consistent with and prepared to provide data able to meet the requirements to demonstrate disease modifying effects from a treatment and able to justify the place of disease mechanism markers as surrogates for clinical efficacy [62]. These latter properties of this methodology are of importance since, by highlighting unique mechanistic advantages for a drug under development, these methods potentially offer routes to FDA accelerated approval of an investigational drug based on novel therapeutic potentials.

Phased Implementation of Drug Development

1. Prior to initiating clinical studies and in parallel to the toxicity and other studies required for human studies two areas of laboratory investigation lay the groundwork for Phases I and II clinical studies:
 - a. *In vitro* kinetic studies of drug-receptor or drug molecular target interactions. These studies estimate important parameters such as required drug concentration at the target; duration of drug activity at the molecular target which will affect initial estimates of dosing intervals to maintain efficacy, to avoid accumulation and adverse effects; and so forth. If anatomical material is available, comparisons should be undertaken between molecular target samples obtained from animals used as models for estimating human dosing and from humans.
 - b. Animal models, for example rats, mice, rabbits, dogs, monkeys, described by the kinetics of drug absorption by routes available for human administration, other routes that may explain characteristics of observed bioavailability and pharmacokinetics. In these models pharmacokinetic and pharmacodynamic estimates establish dose-response relationships and estimate the variability associated with biological differences among and within species. Since drug concentrations and dynamic effects at their target will not in all probability be available in human studies, a particular aim of the animal models is to provide patterns of relationships among indicators that can be measured in humans to allow argument by analogy from human observations to probable target concentrations and effects. Adequate pharmacokinetic and dynamic animal models also indicate the range of bioavailability, rates of elimination, metabolism, and half-lives related to accumulation. Microdosing with labeled compounds can confirm models or,

for compounds where *in vitro* and animal models cannot predict adequately for humans, provide the model needed to proceed to later developmental stages [63,64]. These models provide a background from which human dosing requirements can be initially estimated and against which results in human studies can often usefully be interpreted and understood.

- c. *Ex vivo* drug metabolism studies in human and animal (see b, above) biological matrixes (fractions from blood, liver and potentially target organ) to characterize the time-dependent metabolic profile of the drug and active metabolites to estimate pharmacokinetic and dynamic actions in Phase I and II CTs.
2. To facilitate reaching solutions to problems that inevitably arise as drug development proceeds, sponsors may find useful expanding working relations among the pharmacologists, clinical pharmacologists and clinical researchers involved in pre-clinical and clinical phases of the drug development. For example, predictions from animal models and human data from Phase I may not correspond and simple bench or animal investigations may clarify the issues providing a sounder base for design of Phase II.

For the smaller company with fewer resources to commit to a drug development, the more prevalent practices of separately contracting out to different vendors each of the activities required as steps in development may not be as advantageous as the more integrated and interactive professional relationships just described. Continuity between preclinical and clinical adds resources for problem solving. A functional interaction of Phases I, II, and III, rather than the customary steps from one vendor to the next for each phase, offers the opportunity for the sponsor to overcome, with preparations during Phases I and II, the error variance problems common in Phase III.

3. While most persons in drug development would agree with the concept of designing and executing clinical studies to reach maximum power without excessive Ns, the statistical authority, Joseph Cohen [65], describes the reality of current practices: "We have also lost sight of the fact that the error variance in our observations should challenge us to efforts to reduce it and not simply to thoughtlessly tuck it into the denominator of an F or t test." Cohen advocates against practices that have allowed an N = 1000 to 1500 to become acceptable in AD CTs. Even putting aside the unnecessary tolerance of error that is not addressed directly and corrected, but rather compensated for and buried by a large N, the large N also increases expenses greatly. For example, for a study of N = 1500 costs could increase by a factor of 3 or 4 over costs for an N = 300 to 400. Large Ns increase the burdens of finding and monitoring additional sites. With increased N the intercommunication and close monitoring for quality becomes difficult, if not impossible. Excess numbers of sites, because close monitoring is more difficult, increase the risks of deviations from protocols and reemergence of the problems of error thought to be solved by N.

Important also, but not commonly discussed, are the degrees of selectivity available choosing subjects comparing large and small N studies. Smaller studies can more easily find needed numbers of subjects without coexisting diseases and use of prescribed medications that may produce adverse events able to be confused as due to study drug, and thereby lead to compromises to adequate study drug dosing. Problems recruiting large Ns lead to compromises that are rationalized as 'real world testing conditions' when, in fact, the CT in Phase III aims to confirm for the drug the conditions of use to achieve efficacy and, hopefully, effectiveness for individuals - not to document the shortcomings of the 'real world.' The proposals in this article aim to develop and support with CTs scientific conclusions that describe performance criteria for the 'real world' rather than have drug development suffer from questionable levels

of bias and lacking skills in the ‘real world.’ Once the drug’s potential benefits are established then the problems of achieving those benefits in the real world can be addressed without passing off the consequences of unrelated problems in the study population, such as adverse events unrelated to the study drug leading to inadequate dosing and so forth, onto the drug. Science developed in CTs, not customs in patient care, must set the conditions for drug use if medicine is to become scientific.

“Three factors affect the power of a CT to detect a statistically significant difference between actively treated and placebo treated patients or episodes of patient care: 1.) mean differences among treatment arms, 2.) size of the variance and 3.) number of subjects” [1]. It could be claimed that mean differences among treatment arms can not be increased for a given drug since this property is drug limited. However, Engelhardt and associates’ recent reports document that poorly performing raters reduce the mean differences among treatment arms and, consequently, the power of a statistical test to detect a treatment effect that is actually present but inaccurately reported. Becker’s studies of increased error variance demonstrate how inherent limitations of measurement in methods of observation and rater’s lack of skills, attention, inadequate inquiry of the patient, and so forth contribute to increases in variance, the denominator of statistical tests, reducing power and leading to the ‘tucking’ of the problems into increases in the number of subjects required to detect statistically significant treatment differences that are actually present in research subjects [1,9].

We recommend increased standardizations of assessment methods, monitoring the quality of how raters perform outcome evaluations, and reorientation of the sponsor’s attitudes towards the phases of drug development to capitalize on opportunities that can be created to carry out the increased standardizations and monitoring for quality of outcome evaluations. These operational changes are discussed, below, under 4.

4. The opportunity exists for sponsors to use Phases I and II to test outcome measure administrations for reliability, to recruit sites, and to train, evaluate, and screen site evaluators for participation in Phase III. To achieve error control in clinical trials, during Phases I, II, and III we recommend the following changes in drug development clinical phase management practices:

a. Standardization of outcome evaluations

- i. Evaluators administer all outcome measures using the same administration protocol for the outcome scale (see for example Williams [66]).
- ii. Selected evaluation sessions conducted by each evaluator are taped and the tapes reviewed for quality by a professional expert in the use of the outcome measure.
- iii. This expert’s rating of the evaluator’s performance uses a standard evaluation form that addresses rater adherence to the protocol, use of follow-up questions, adequacy of the rater’s clarification of ambiguous patient responses, the neutrality of the rater’s attitude, rapport established with the subject, adequacy of information developed in relation to the requirements of the item being evaluated, and the duration of the interview (see for example, Lipsitz *et al.* [67] who provide a scale for use with the Hamilton depression scale and the use of this scale in Engelhardt [2]).
- iv. Rating of some subjects by expert raters to detect in the site evaluator’s ratings the presence of bias, failed accuracy, especially failures to respond appropriately to changes in the clinical state of the subject, to use as a ‘Gold Standard’ for sites’ performances, and to use as part of the data analysis.

- v. Consideration by the sponsor of having study ratings done by expert raters not associated with the research sites and to limit the sites to management of the patient under the Study Protocol.
- vi. Use of a clinically well-experienced, well-trained raters and consideration of using different raters at each evaluation to improve discrimination of changes associated with active treatment [26,33,53].

b. Use of Phase I and II to train, evaluate, screen, and qualify for Phase III participation rating evaluators and sites

- i. Research sites are recruited to perform clinical evaluations of subjects for each of the Clinical Phases, I, II, and III. Thus potential Phase III evaluators can be trained prior to and evaluated in each of Phases I and II. The data acquisition steps required in each Phase, I, II, and III, serve as natural opportunities to develop quality in the data collected in later phases since raters who can not demonstrate the needed skills in Phases I and II can be excluded from Phase III.
- ii. Evaluators administer in each Phase outcome measures under the conditions described above in 3. a. The evaluator's abilities to provide evaluations consistent with expert evaluations and compliant with the protocol for administration determine whether or not the evaluator is retained for further training and supervision in the next phase or removed from participation. Sponsors can manage enlistment of sites over the whole clinical period, Phases I, II, and III, thus developing working relations with and stability for sites, training raters and testing abilities to monitor performance such that variance, and therefore N, costs and duration, are reduced for Phase III, and enlisting subjects for Phases prior to phases. This latter relation with subjects provides the opportunity to develop longitudinal data about each subject's disease allowing double-blind and open exposures to drug to be analyzed in relation to long-term course and outcome.

CONCLUSIONS

Current drug development practices tolerate fragmentation of investigations that overlook how work with the human species is, in the perspective of a well investigated compound, an extension of work *in vitro* and in other than human species. In our view, the major problems in AD drug development are expressed by investigator's insufficient attention to dosing parameters and, for compounds whose clinical effects can be measured only with rating scales, lack of awareness of how imprecision and inaccuracy expand variance in outcome measurements. Drug development is pulled in many different directions as investigators struggle for funding, to find and maintain subjects in CTs, attempt to accelerate into clinical phases with the cost that human responses can not be evaluated in the light of well characterized *in vitro* and animal models of drug actions, and so forth.

Our view is that greater recognition must also be given to how drugs act at molecular targets and, in turn, to characterizing how modifications in pathology at molecular targets can be measured for use as outcome variables and their relations to clinical outcomes in patients. Certainly, molecular measures may not yet be acceptable indicators for outcomes of either drug efficacy or disease progression by the US FDA, but they may nevertheless prove of value to target therapeutic windows, optimize interpretations of clinical outcome measures and to avoid adverse actions. AD is currently at best characterized as a multi-dysfunctional molecular condition with possibly interrelated and possibly independent pathologies in amyloid formation, tau abnormalities, cerebro-vascular amyloid accumulations, and loss of Nucleus Basalis acetylcholine modulation of cortical neuro-transmission. Currently available drugs and

drugs under development for AD each target one of these mechanisms, suggesting the need for a combination of symptomatic or disease modifying treatments as a next step. CTs that evaluate treatment effects against one or more of these molecular targets potentially offer more rational clinical grounding for combination therapies in AD and experimentally evidence the contributions that specific mechanisms make to AD behavioral pathology. With the advent of new resources to evaluate AD patients—structural MRI, molecular neuroimaging with PET, cerebrospinal fluid analyses, and so forth—molecular mechanisms of AD become more accessible as targets for treatment interventions [36,68].

The difficulties we identify as interfering with the successful conduct of drug development and CTs also raise two compromises to medical ethics that have practical negative consequences for drug development. First, the Nuremberg Code [69] in its second requirement states “The experiment should be such as to yield fruitful results for the good of society...” Under current practices investigators and commercial sponsors can keep secret results from human research even though this practice has been acknowledged as “scientific misconduct” [49]. Whether the sciences of dose finding could benefit from the commercial studies of compounds not reported in the literature can not be evaluated because business interests to hold human research results as secrets prevail in law over the Nuremberg Code [69]. The full Nuremberg Code does not have status as international, national or regional law in any jurisdictions. The World Medical Association notes that negative studies ‘should’ be published rather than requiring public reporting and access as a matter of medical and scientific ethics. The scientific misconduct associated with not providing public access to results from human investigations and experimentation is not open to legal challenge and thus can be expected to continue except when investigators or sponsors elect to do otherwise.

The Nuremberg Code [69], incorporated into the Declaration of Helsinki and adopted by the World Medical Association, states “The experiment should be conducted only by scientifically qualified persons.” The data from Engelhardt and colleagues, Becker, Targum and others suggest that many professionals recruited as evaluators in CTs are not skilled sufficiently to qualify for participation in scientific clinical research. In view of the increasing competition for clinical sites for CTs and for patients as subjects, we can not expect commercial sponsors to risk alienating potential research sites by becoming more selective; yet, this appears to be exactly what drug development requires to overcome the problems of bias, inaccuracy and imprecision that have led to failed CTs and CTs in which placebo group behaviors do not model the established concept of AD as a deteriorating neurological condition. AD CTs, like many other neurological CTs, may involve sites, such as eastern Europe and elsewhere, where AD demographics and the levels of health care may potentially be less well characterized, investigators less skilled and experienced in diagnosing, assessing, and managing AD, and clinical outcome measures are more susceptible to cultural and language differences. As Cohen [65] notes, investigators can bury variance in the denominator of statistical tests by adding Ns to CTs but only with negative consequences. Recently, Cogger [4] demonstrated the impact on effect sizes from burying variance due to inaccuracy, failures of raters to follow rating protocols, inattention, and bias on the part of clinical raters. Following Kobak *et al.* [34] and Demitrack [32], drug development sponsors must be prepared to not use investigators who can not demonstrate the required levels of skills.

Large Ns increase the exposure of patients to placebos and investigational compounds, delay evaluations of new compounds in AD because of the increased time to completion of CTs, and drug development time and costs greatly expand with the need to recruit and monitor sites. These pressures lead sponsors to enlistment of sites with personnel not scientifically skilled to provide the required precision and accuracy in observations, require planners of CTs to loosen criteria in order to complete recruitment, and lead to long recruitment times compared to more tightly designed and controlled studies not dependent on numbers recruited to develop the

required power to test the hypothesis of the study. Insufficient *in vitro* and animal modeling contribute to investigators inadequately characterizing dosing parameters prior to Phase III CTs. The ability of investigators to not publish and keep from public scrutiny negative results or even positive results considered ‘business secrets’ causes mistakes in drug development to be repeated and patients to be exposed to treatment conditions that do not provide maximum available benefits. Current problems with human errors in clinical trials make it difficult to differentiate drugs that fail to evidence efficacy from apparent failures due to Type II errors. AD drug development faces challenges that must be addressed if drug development is to become more effective, more economical, streamlined to provide more rapid feedback to medicinal chemists, pharmacologists and numerous other bench scientists responsible for new compounds, more fully grounded in science and investigators are to avoid scientific misconduct.

Acknowledgements

The authors are supported in part by the Intramural Research Program of the National Institute on Aging, NIH. The views expressed within this article are those of the authors and may not represent those of the National Institute on Aging, NIH.

References

1. Becker RE. Lessons from Darwin: 21st Century designs for clinical trials. *Curr Alzheimer Res* 2007;4:458–467. [PubMed: 17908050]
2. Engelhardt N, Feiger AD, Cogger KO, Sikich D, DeBrotta DJ, Lipsitz JD, et al. Rating the Raters: Assessing the Quality of Hamilton Rating Scale for Depression Clinical Interviews in Two Industry-sponsored Clinical Drug Trials. *J Clin Psychopharmacol* 2006;26:71–74. [PubMed: 16415710]
3. Kobak, KA.; DeBrotta, DJ.; Engelhardt, N.; Williams, JEW. Site vs. centralized raters in a clinical depression trial. National Institute of Mental Health, New Clinical Drug Evaluation Unit. 46th Annual Meeting; Boca Raton, FL. 2006.
4. Cogger KO. Rating rater improvement: a method for estimating increased effect size and reduction of clinical trial costs. *J Clin Psychopharmacol* 2007;27:418–20. [PubMed: 17632240]
5. Frampton M, Harvey RJ, Kirchner V. Propentofyllin for dementia. *Cochrane Data Base Syst Rev* 2003;2:CD002853.
6. Becker RE, Giacobini E. Mechanisms of cholinesterase inhibition in senile dementia of the Alzheimer type: clinical, pharmacological, and therapeutic aspects. *Drug Dev Res* 1988;12:163–195.
7. Becker, RE.; Giacobini, E. Advances in the therapy of Alzheimer’s disease: issues in the study and development of cholinomimetic therapies. In: Giacobini, E.; Becker, R., editors. *Current Research in Alzheimer Therapy*. Taylor & Francis; New York, NY: 1990.
8. Becker, RE.; Moriearty, P.; Unni, L. The second generation of cholinesterase inhibitors: clinical and pharmacological effects. In: Becker, R.; Giacobini, E., editors. *Cholinergic Basis for Alzheimer Therapy*. Birkhäuser; Boston: 1991. p. 263-296.
9. Becker RE. Lessons from Darwin: Evolutionary biology implications for Alzheimer’s disease research and patient care. *Curr Alzheimer Res* 2005;2:319–326. [PubMed: 15974897]
10. Greig NH, Sambamurti K, Yu QS, Brossi A, Bruinsma GB, Lahiri DK. An overview of phenserine tartrate, a novel acetylcholinesterase inhibitor for the treatment of Alzheimer’s disease. *Curr Alzheimer Res* 2005;2:281–90. [PubMed: 15974893]
11. Greig NH, Ruckle J, Comer P, Brownell L, Hollo way HW, Flanagan DR Jr, Canfield CJ, Burford RG. Anticholinesterase and pharmacokinetic profile of phenserine in healthy elderly human subjects. *Curr Alzheimer Res* 2005;2:483–92. [PubMed: 16248851]
12. Giacobini, E. Cholinesterase inhibitors: from the Calabar bean to Alzheimer therapy. In: Giacobini, E., editor. *Cholinesterases and Cholinesterase Inhibitors*. Martin Dunitz; London: 2000. p. 181-226.
13. Birks J. Cholinesterase inhibitors for Alzheimer’s Disease. *Cochrane Database Syst Rev* 2006;1:CD005593. [PubMed: 16437532]
14. Klein J. Phenserine. *Expert Opin Investig Drugs* 2007;16:1087–97.

15. Thatte U. Phenserine Axonyx. *Curr Opin Investig Drugs* 2005;6:729–39.
16. López-Arrieta JM, Schneider L. Metrifonate for Alzheimer's disease. *Cochrane Database Syst Rev* 2006;2:CD003155. [PubMed: 16625573]
17. Unni LK, Radcliffe J, Latham G, Sunderland T, Martinez R, Potter W, et al. Oral administration of heptylphysostigmine in healthy volunteers: a preliminary study. *Meth Find Exp Clin Pharmacol* 1994;16:373–376.
18. Moreiarty P, Becker RE. Inhibition of human brain and RBC acetylcholinesterase by heptylphysostigmine. *Meth Find Exp Clin Pharmacol* 1992;14:615–621.
19. Unni LK, Becker RE. Determination of heptylphysostigmine in plasma by high-performance liquid chromatographic with electrochemical detection. *J Chromatogr Biomed Appl* 1992;573:275–281.
20. Giacobini E, Becker R. One hundred years after the discovery of Alzheimer's disease: a turning point for therapy? *J Alzheimers Dis* 2007;12:37–52. [PubMed: 17851193]
21. Becker RE, Colliver JA, Markwell SJ, Moriearty PL, Unni LK, Vicari S. A Double-Blind, Placebo-Controlled Study of Metrifonate, an Acetylcholinesterase Inhibitor, for Alzheimer's Disease. *Alzheimer's Dis Assoc Disord* 1996;10:124–131.
22. Becker RE, Colliver JA, Markwell SJ, Moriearty P, Unni LK, Varney A, et al. A Six Month Study of the Effects of Metrifonate on Cognitive Decline in Alzheimer's Disease. *Alzheimer's Dis Assoc Disord* 1998;12:54–57.
23. Becker RE, Markwell S. Problems arising from generalizing of treatment efficacy from clinical trials in Alzheimer's disease. *Clin Drug Invest* 2000;19:33–41.
24. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol Psychiatry* 2000;47:762–766. [PubMed: 10773186]
25. Lachin JM. The role of measurement reliability in clinical trials. *Clin Trials* 2004;1:553–566. [PubMed: 16279296]
26. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol* 2007;27:1–5. [PubMed: 17224705]
27. Woods SW, Gueorguieva RV, Baker CB, Makuch RW. Control group bias in randomized atypical antipsychotic medication trials for schizophrenia. *Arch Gen Psychiatry* 2005;62:961–970. [PubMed: 16143728]
28. Kahn A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol* 2002;22:40–45. [PubMed: 11799341]
29. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry* 1995;52:867–871. [PubMed: 7575107]
30. These ME. Studying new antidepressants: if there were a light at the end of the tunnel, could we see it? *J Clin Psychiatry* 2002;63(Suppl 2):24–28.
31. Muller MJ, Szegedi A. Effects of interrater reliability of psychopathologic assessment on power and sample size calculations in clinical trials. *J Clin Psychopharmacol* 2002;22:318–325. [PubMed: 12006903]
32. Demitrack MA, Paries D, Herrera JM, DeBrotta D, Potter WZ. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull* 1998;34:19–24. [PubMed: 9564194]
33. Rickels K, Robinson DS. Why do clinical trials fail? *J Clin Psychopharmacol* 2007;27:419–420.
34. Kobak KA, Lipsitz JD, Williams JEW, Engelhardt N, Bellew KM. A new approach to rater training and certification in a multicenter clinical trial. *J Clin Psychopharmacol* 2005;25:407–412. [PubMed: 16160614]
35. Greig NH, De Micheli E, Holloway HW, Yu QS, Utsuki T, Perry TA, et al. The experimental Alzheimer drug phenserine: preclinical pharmacokinetics and pharmacodynamics. *Acta Neurol Scand* 2000;176:74–84.
36. Kadir, A.; Andreasen, N.; Almkvist, O.; Wall, A.; Forsberg, A.; Engler, H., et al. Effect of phenserine treatment on brain functional activity and amyloid in AD patients (submitted). *Neurodegen Dis*; 8th International Conference on Alzheimer's disease/Parkinson's Disease Salzburg; Austria. March 14–18, 2007; 2007. p. 316-7.

37. Axonyx, Me. SEC filed press release of March 11, 2005 (Results of phase III clinical trial for phenserine in Alzheimer's disease) <http://www.secinfo.com/d13Wqv.zlj4.d.htm>
38. Winblad, B.; Giacobini, E.; Froelich, L.; Bruinsma, G.; Walters, E.; Friedhoff, L. Double-blind placebo-controlled evaluation of the safety and efficacy of phenserine tartrate for the treatment of mild to moderate Alzheimer disease. 9th International Geneva/Springfield Symposium on Advances in Alzheimer Therapy. Abstract 11 A; April 19–22; 2006.
39. Solomon B. Clinical immunologic approaches for the treatment of Alzheimer's disease. *Expert Opin Investig Drugs* 2007;16:819–28.
40. Feldman HH, Ferris S, Winblad B, Sfikas N, Mancione L, He Y, et al. Effect of rivastigmine on delay to diagnosis of Alzheimer's disease from mild cognitive impairment: the InDDEX study. *Lancet Neurol* 2007;6:501–512. [PubMed: 17509485]
41. Petersen RC. MCI treatment trials: failure or not? *Lancet Neurol* 2007;6:473–475. [PubMed: 17509475]
42. Blass JP, Cyrus PA, Bieber F, Gulanski B. The Metrifonate Study Group. Randomized, Double-Blind, Placebo-Controlled, Multicenter Study to Evaluate the Safety and Tolerability of Metrifonate in Patients with Probable Alzheimer Disease. *Alzheimer Dis Assoc Disord* 2000;14:39–45. [PubMed: 10718203]
43. Ingenta. Clinical trials with metrifonate have been stopped by Bayer. *Reactions* 1998;1:4–4(1).
44. Becker, RE.; Moriearty, P.; Surbeck, R.; Unni, L.; Varney, A.; Vicari, S. Second and third generation cholinesterase inhibitors: clinical aspects. In: Giacobini, E.; Becker, R., editors. *Alzheimer Disease: Therapeutic Strategies*. Birkhauser; Boston: 1994. p. 172–178.
45. Ringman JM, Cummings JL. Metrifonate: update on a new antidementia agent. *J Clin Psychiatry* 1999;60:776–82. [PubMed: 10584768]
46. Imbimbo B. Pharmacodynamic-tolerability relationships of cholinesterase inhibitors for Alzheimer's disease. *CNS Drugs* 2001;15:375–90. [PubMed: 11475943]
47. Becker RE, Unni L. Are there selective effects of AChEIs in AD? 1998Unpublished
48. Chalmers I. Underreporting research is scientific misconduct. *JAMA* 1990;9:1405–1408. [PubMed: 2304220]
49. Becker R, Colliver J, Elble R, Feldman E, Giacobini E, Kumar V, et al. Effects of Metifonate, A Long Acting Cholinesterase Inhibitor in Alzheimer Disease: Report of an Open Trial. *Drug Dev Res* 1990;9:425–434.
50. Jeglic E, Kobak KA, Engelhardt N, Williams JEW, Lipsitz JD, Salvucci D, et al. A novel approach to rater training and certification in multinational trials. *Int Clin Psychopharmacol* 2007;22:187–191. [PubMed: 17519640]
51. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet-based technologies: a comparison to traditional rater training. *J Psychiatr Res* 2006;3:192–199. [PubMed: 16197959]
52. Kobak KA, Lipsitz JD, Feiger A. Development of a standardized training program for the Hamilton Depression Scale using internet-based technologies: results from a pilot study. *J Psychiatr Res* 2003;37(6):509–15. [PubMed: 14563382]
53. Targum SD. Evaluating rater competency for CNS clinical trials. *J Clin Psychopharmacol* 2006;26:308–310. [PubMed: 16702896]
54. Fearnley JM, Lees AJ. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain* 1991;114:2283–2301. [PubMed: 1933245]
55. Arendt T, Bigl V, Arendt A, Tennstedt A. Loss of neurons in the nucleus basalis of Meynert in Alzheimer's disease, paralysis agitans and Korsakoffs disease. *Acta Neuropathologica* 1983;61:101–108. [PubMed: 6637393]
56. Clark CM, Sheppard L, Fillenbaum GG, Galasko D, Morris JC, Koss E, et al. Variability in annual Mini-Mental State Examination score in patients with probable Alzheimer disease: a clinical perspective of data from the Consortium to Establish a Registry for Alzheimer's Disease. *Arch Neurol* 1999;56:857–862. [PubMed: 10404988]
57. Feldman H, Gauthier S, Hecker J, Vellas B, Subbiah P, Whalen E, Donepezil MSAD Study Investigators Group. A 24-week, randomized, double-blind study of donepezil in moderate to severe Alzheimer's disease. *Neurology* 2001;57:613–620. [PubMed: 11524468]

58. Becker RE. Modifying clinical trial designs to test treatments for clinical significance in individual patients. *Clin Drug Invest* 2001;21:727–733.
59. Becker RE. Focusing pharmaceutical research on patient care: using clinical trials to develop clinical decision rules. *Clin Drug Invest* 2002;22:269–278.
60. Becker RE. Treating Alzheimer's disease: Do clinical trial methods produce unnecessary practitioner dilemmas? *Am J Alzheimer's Dis Other Dementias* 2006;21:81–84.
61. Donohue JM, Cevasco M, Rosenthal MB. A decade of direct-to-consumer advertising of prescription drugs. *N Engl J Med* 2007;357:673–8. [PubMed: 17699817]
62. Mani RB. The evaluation of disease modifying therapies in Alzheimer's disease: a regulatory viewpoint. *Statist Med* 2004;23:305–314.
63. Bauer M, Langer O, dal-Bianco P, Karch R, Brunner M, Abraham A, et al. A positron emission tomography microdosing study with a potential anti-amyloid drug in healthy volunteers and patients with Alzheimer's disease. *Clin Pharmacol Ther* 2006;80:216–27. [PubMed: 16952488]
64. Lappin G, Kuhn W, Jochemsen R, Kneer J, Chaudhary A, Oosterhuis B, et al. Use of microdosing to predict pharmacokinetics at the therapeutic dose: Experience with 5 drugs. *Clin Pharmacol Ther* 2006;80:203–215. [PubMed: 16952487]
65. Cohen J. Things I Have Learned (So Far). *Am Psychol* 1990;45:1304–1312.
66. Williams JEW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry* 1988;45:742–747. [PubMed: 3395203]
67. Lipsitz JD, Kobak KA, Feiger A, Sikich D, Maroz G, Engelhard A. The Rater Applied Performance Scale (RAPS); development and reliability. *Psychiatry Res* 2004;127:147–151. [PubMed: 15261713]
68. Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol* 2007;6:734–746. [PubMed: 17616482]
69. Nuremberg Code. Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No 10: Nuremberg October 1946–April 1949. 2. Washington: U.S. Government Printing Office; 1949. Permissible Medical Experiments; p. 181–182.

Fig. (1).
A Model Flow for a Representative AD Drug Development

Fig. (2).

A Model Time-Flow for Drug Development: Ambitious; Yet, Possible