Alzheimer's Disease Drug Development: Old Problems Require New Priorities

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: Alzheimer's disease (AD) clinical drug development and patient care depend on rating instruments, research designs and methods, and translations of clinical trial (CT) results into the clinic without support from standardized protocols able to control (i) random measurement error intrusions into observations, (ii) inaccuracy and bias introduced by clinical evaluators, (iii) conformity of research sites to conditions of research protocols, (iv) the ability of the CT to model for practitioners the most effective use of the drug with individual patients, and (v) other factors able to invalidate research and patient care data. This relaxed attitude with regard to AD methods may be changing now with Alzheimer's Disease Neuroimaging Initiative (ADNI) evidence that carefully standardized protocols are needed to validate biomarkers for use in AD diagnosis, drug development, and patient care.

In the fields of psychiatry and AD, recent studies have detected serious inaccuracies, imprecision, biases and compromises of study protocols able to invalidate CT outcome data and conclusions drawn from these data. This limited but troubling evidence reinforces ADNI calls for more detailed methodological protocols. Based on the limits to precision and accuracy associated with rated outcomes in CTs and patient care, we call for priority to be given to the qualification and use of biomarkers as outcome variables in AD drug development and patient care and, to insure effective uses of biomarkers, to development of protocol guided practices being modeled in ADNI research. To meet clinical pharmacology's therapeutic aims we conclude that AD CTs need to set for clinicians the conditions of use of drugs shown efficacious, biomarker surrogate endpoints as drug targets, and not to function merely as tests for efficacy conducted under field conditions determined by current clinical practices.

Keywords: Clinical trials, drug development, Alzheimer's disease, measurement errors, biomarkers, protocols.

INTRODUCTION

Dubois *et al*. [1] conclude that advances in genetics, imaging, biochemistry, and clinical descriptions necessitate revisions of criteria for diagnosing Alzheimer's disease (AD). These recent advances in AD relevant sciences and ongoing difficulties with clinical pharmacology methodologies potentially affect AD drug development [2-4]. Pangalos *et al*. [2] point out that success rates with AD drugs remain below the already low 7% rate for all central nervous system drugs. The high failure rate in development recently led us to consider whether AD drug developments fail due to current methods and practices and not solely from candidate drugs lacking efficacy or safety [3, 4].

As an initial test for possible design or execution flaws affecting AD clinical trials (CT) we analyzed in detail CTs for two failed drugs strongly supported by preclinical evidence and by CT proven efficacy for other drugs in their class [5]. Our studies of these failed trials suggested that methodological flaws in AD clinical pharmacology practices contributed to the drug failures and that these flaws, prepared to undermine other drug developments, may lurk undetected within current AD drug development practices [3]. In Becker and Greig [3] we attempted to address a range of these factors (Table **1**).

We then randomly selected 10 AD drug candidates from each of four groups: approved AD

*Address correspondence to this author at the Drug Design & Development Section, Laboratory of Neurosciences, Intramural Research Program, National Institute on Aging, National Institutes of Health, Baltimore, MD 21224, USA; E-mail: rebecker2008@comcast.net

1871-5273/08 $55.00+.00

Table 1. Already Acknowledged Threats to AD CT Validity

- Inadequate preparations to effectively dose drugs in Phase III
- Unreliability of outcome measures
- Insufficient care implementing and monitoring quality controls at research sites
- Insufficient development in CTs of resources to enable practitioners to optimize uses of CT evidence in patient care
- Insufficient disease and drug effects modeling in preclinical and early clinical stages
  - o to support choices of research designs
  - o to provide always clear interpretations of outcomes in Phase III CTs
- Inadequate consideration of whether repeated drug failures undermine current animal models as predictors.

drugs; drugs in development; failed or abandoned drug candidates; and mild cognitive impairment (MCI) drugs. We developed a 56 item list of methodological queries aimed to determine whether investigators address sources of error already identified as putting CTs at risks for failure. The queries covered six areas of concern: 1) drug, mechanisms, pharmacological activities, design, and publication; 2) dosing; 3) research subjects; 4) outcome measures; 5) research sites and investigators, and; 6) protocols to control methods and transfer of findings to clinical practice [3]. On the whole, we found investigators reporting little attention to almost all potential problem areas. We estimated that, overall, 76% of negative AD drug developments went unreported, which confirms earlier analysis [6]. We found additional support for our earlier proposals that problems of variance and its consequences are not adequately addressed in planning for CTs and will remain problematic so long as CTs remain dependent on clinically rated outcome measures [3, 4]. In this paper we review evidence in support of the view that many of the most recalcitrant threats to successful AD drug development devolve from the current dependence of AD clinical investigations on clinical raters. Based on inherent limitations to reliability and validity due to human observational errors, we argue the need to more strongly focus AD research efforts on developing and using biomarkers as surrogate endpoints in AD CTs and clinical practice.

BACKGROUND

When Clinical Pharmacology Practices Catch a Cold, AD Drug Developments Get Pneumonia

Pangalos and colleagues [2] recently reviewed early stages of drug development for factors leading to drug attrition and underestimation of clinical importance in successful CTs. We and others found similar risks from practices in clinical phases [3-5, 7-10]. Authors appear to agree that under current conditions of drug development both drugs can fail and investigators can fail drugs [4, 5]. In Table **1** we list sources of error we and others have identified as putting AD CTs and drug developments at risk of failing. Our concerns with these risks are that Type II errors may eliminate effective drugs from development and that CTs will continue to fail to specify conditions of drug use practitioners require to optimize drug effectiveness in clinical practice.

Core Problems in AD Clinical Pharmacology and their Consequences

Two sources of measurement errors, imprecision and inaccuracy, have been shown to undermine the integrity of neuropsychiatric CTs [4, 5,7,8]. Imprecision is the variation or random measurement error that can not be avoided in repeated applications of a test or rating scale. Inaccuracy is the failure of a test or rating to reflect the state of affairs being measured. In a comparison of CT outcomes for two cholinesterase inhibitors (ChEI), Becker [5] showed how imprecision in one CT, by increasing variance, reduced power and could account for the failure of the CT. Engelhardt *et al*. [7] used Williams' protocol for administration of the Hamilton Depression Scale (HDS) [10] to allow experts well skilled with this scale to evaluate the practices of raters in two antidepressant trials. They found that over 50% of raters, although previously trained in HDS administration, could not detect active drug effects. Raters who failed to detect changes in active drug treated patients were much less compliant with Williams' protocol. Cogger [8] determined that the raters who were non-compliant with these protocol conditions accounted for a 64% reduction in effect sizes in the studies. He estimated that using only raters capable of accuracy could have reduced sample sizes by 87%.

Similar evidence supporting the importance of protocols and of investigators adhering to CT protocols motivated the Alzheimer's Disease Neuroimaging Inititative (ADNI) to standardize "ADNI Methods" now used for imaging at 57 sites and in CTs [11-15]. Preliminary evidence developed by the ADNI found variations between makers of imaging equipment, models, software, and methods to collect and assay biomarker samples with consequential effects on the accuracy of data [11-15].

Methodological protocols aim to bring practices into greater compliance with good practice performance standards; however, certain levels of errors are inherent in scales, tests, and instruments and all forms of error and biases will be magnified by unskilled or non-compliant users. For example, Becker and Markwell [16] showed that well trained and highly experienced raters using the Mini-Mental State Examination and Alzheimer Disease Assessment Scale-Cognitive Subsection encountered levels of imprecision that exceeded the average drug effect expected with ChEIs and the average AD patient's yearly decline. The same scales in less experienced hands at non-specialized sites resulted in much larger variance [5].

Training of raters to provide accuracy and precision across sites, CTs, and raters, shows mixed results. Demitrack [17] found that some raters could not improve performances to required levels regardless of training while other investigators found improvements with training in even the most experienced raters [18]. At present, errors inherent in rated AD observations appear to limit the utility of rating scales in spite of high reported reliabilities [7, 16].

Somewhat analogous validity-limiting problems for CTs arise from excessive heterogeneity of CT subject samples, from volunteer subject characteristics not matching groups eligible for a CT but who refuse to participate, and from other difficulties establishing external validity for AD CTs. Kobak *et al*. [9] found site raters in a study of antidepressants exaggerating pathology presumably to fill CT subject quotas. Rachetti *et al*. [13] and Visser *et al*. [14] show how diagnostic criteria varied considerably among studies of MCI and did not select the same subjects when compared. DuBois *et al*. [1] emphasize inclusion of biomarkers in AD criteria to control this sample heterogeneity. Given the heterogeneity of samples, their selection by convenience, not randomly, and risks from measurement errors for the validity of observations needed to apply entry criteria in AD research, practitioners can easily mistake whether or not a patient matches on essential features with CT subjects.

Each of these problems, potentially limiting validity, cannot be overcome without improvements in the outcome instruments used with AD patients. As we present in Table **2**, imprecision easily couples with rater inaccuracy. Alone or in combination these factors reduce the power of the proposed CT to detect a difference that is present between treatment conditions. Customary practices are to increase the number of subjects in the CT to reach required power. These increased numbers of subjects require multiple research sites. The addition of sites increases the needs for training and monitoring investigators and raters to insure compliance with study protocols. Often, because of competition among studies for sites, sponsors will use sites with raters not experienced with AD patients and not trained and familiar with the instruments used as outcome measures in the CT [7, 18, 19]. As Kobak *et al*. [9] detect, pressures to recruit patients can lead sponsors to relax protocol conditions to allow more candidates to qualify for the study or to sites cutting corners or using biased ratings to meet quotas.

We are concerned that core and derivative problems with AD CTs flow or cascade one to another. The witness from Principal Investigators is that this cascade-measurement errors leading to compromised power leading to increased numbers of subjects leading to multiple sites leading to increased intersite variance-progressively invalidates CTs faster than even conscientious sponsors and investigators can intervene to correct misdirections of efforts [19]. Given that inherent error sources in outcome instruments and in human users initiate this cascade of error sources, we suggest that AD clinical pharmacology needs to turn its attention to developing and validating outcome measures with error components that will not launch cascades of negative consequences for validity.

Table 2. Comparisons of Methodological Problems in CTs Using Clinician Rated Outcome Measures and Surrogate Biomarker Endpoints

I. Core Problems	Influence with Clinician Rated Outcomes	Influence with Surrogate Endpoints
Rating Instrument Imprecision and Inaccuracy	High	Low
Clinical Rater Imprecision and Inaccuracy	High	Low
Rater Observer Bias	High	Low
Failure to Control Practices with Protocols	High	Needed Consistent with Good Laboratory Practices
Sampling Bias Excess Heterogeneity Non-Representative Volunteers	 High High	Possible; however, Screening with Biomarker Allows Molecular Targeting and Reduces Bias Effects
External Validity and Generalizability	Low	High Generalize to Patients with Targeted Biomarker
II. Derivative Problems		
Increased Variance in Data Sets from Core Problems	High	Low
Reduced Power of CTs from Increased Variance	High	Low
Increased Numbers of Subjects to Provide Power to CT Analysis	High	Low
Increased Research Sites and Raters to Provide Subjects	High	Low
Increased Needs for Training and Monitoring	High	Low
Pressures to Compromise Protocols Sponsor Relaxation of Entry Criteria or Methods to Aid Recruitment Increased Pressures on Sites for Subjects	High High High	Low Low Low
Increased Sampling Bias From Clinically Unskilled Raters Large N increases risk of non-compliance with protocols	 High High	 Low Low
Placebo Group Improvements Unskilled Observer Bias Non-Specific	 High High	 Not with Laboratory Outcomes Not Relevant
Problems of Matching Patients in Clinical Care to CT Samples	High	Aided by Screening Using Biomarkers

OPPORTUNITIES FOR IMPROVED RELIABILITY AND VALIDITY IN AD CTS AND CLINICAL PRACTICES

Magnetic resonance imaging (MRI), positron emission tomography (PET), single positron emission computerized tomography (SPECT), and spinal fluid biochemistry now offer glimpses into the progression of pathologies present as AD patients deteriorate in cognition and behaviors. These biomarkers do not entirely escape measurement errors and other problems that currently compromise AD CTs [20, 21]. On the other hand, bioassays, done with standardized techniques under good laboratory practices (GLP) [22, 23] offer AD researchers opportunities to improve accuracy and precision of measurements to levels adequate to improve diagnoses, to potentially develop molecular targets for drug interventions, to reduce the number of subjects required in AD CTs [24, 25], and to manage patients in clinical practice with reduced risks of measurement error [5, 26]. Biomarkers offer CT investigators this promise, in part, because of the increased accuracy and precision as values obtained under GLP conditions replace human judgments.

Any scientist understands that precision, accuracy, and freedom from bias are hard-won conditions requiring, at each step, from collection of samples to analysis of data, careful preliminary experimentation, demonstrations that values are free from interfering errors, ongoing monitoring of quality with controls, trained and experienced personnel, and so forth. Neuroimaging, biochemical assays, and other laboratory aided assessments do not automatically overcome the problems that currently follow from CTs being overly dependent on clinical ratings. Current evidence suggests that these quantitative biomarker measures can potentially offer accuracy and precision not available with clinical ratings [24]. Unfortunately, reliabilities, the statistics commonly used to qualify clinical rating scales, may not reflect levels of imprecision or inaccuracy that can invalidate AD CTs. It remains possible that improved methods of clinical rating may become available; however, existing studies suggest that reliability coefficients will not suffice to quantify effects from inaccuracy and imprecision. Given the problems with inaccuracies, imprecision, biases, and protocol deviations detected in studies that have been carried out using psychiatric and AD rating scales, we are not sanguine that clinical ratings will meet the needs of AD researchers and clinicians.

At present, a range of neuroimaging and biochemical assays tempt AD investigators with the promise that surrogate endpoints can be developed to support diagnosis and to measure disease severity in AD [27]. Currently, biomarkers do not always provide advantages over existing clinical assessments [28]. For example, in a four year study Feldman *et al.* [29] found behavior, apathy, cognitive executive functions, attention, and verbal fluency, but not baseline MRI assessments, predicting progression to AD from MCI. Even though the hippocampus has been regarded as an early and prominent site of atrophy in AD, MRI whole brain and ventricular, but not hippocampal, volume changes, were related to progression in the Feldman *et al.* study. On the other hand, biomarkers may not yet have received fair evaluations because of study conditions. Rachetti *et al.* [13] and Vissar *et al.* [14] found MCI criteria not selecting consistent subject samples across studies. Vallas *et al.* [25] observe that biomarkers have not yet been used to target patient subgroups with pathologies against which, in animal models, the drug candidate has shown promise. Aisen and Vallas find protocol violations interfering in multicenter trials [19]. Unfortunately, for the foreseeable future, some authors see preclinical and clinical phase methodological and practice shortcomings seriously impairing systematic developments of biomarkers with specific utilities in pre-AD and AD [30].

Negative implications from unreliability in CT outcome measures are acknowledged [30, 31]. Experiences in various fields of medicine show that methodological advances in research and patient care depend on development of standardized protocols for using instruments, tests, and assays [11, 30-32]. If methodological standardization is the problem in AD that it is thought to be by some authors, then ADNI efforts to standardize AD biomarker methodologies and develop protocols able to systematize and control applications of biomarkers may be crucial to increasing the yield from AD drug developments [11, 24].

Systematic and controlled methods for measuring biomarker indicators of brain pathologies potentially can avoid the cascade of risks for current AD CTs. More accurate and precise values for outcome measures should support earlier diagnoses, facilitate CTs identifying drugs for disease modifying effects, and better control targeted therapeutic interventions in CTs and clinical practice. Today, imaging and biochemical biomarkers potentially offer measurements to AD investigators with greatly improved precision and accuracies and better insulation from effects due to site laxness and carelessness.

Pangalos *et al.* [2] recommend "sensitive efficacy" markers of effectiveness of therapeutic interventions as needed for improved central nervous system drug discovery and development. To indicate efficacy, biomarkers must provide sufficiently reliable, accurate and precise, quantitative indications of the severity of pathological processes in AD and predict long-term patient benefits. Currently practiced PET monitored microdosing demonstrates one practical application of quantitative biomarkers and their utility [33]. This technology provides indications of brain target site concentrations of a drug, the pharmacokinetics associated with the target concentrations, and models for CT dosing of an investigational compound.

Unfortunately, although much research is promising, not all authors are sanguine that biomarkers, once characterized, can be rapidly and strongly linked to clinical outcomes. The Food and Drug Administration (FDA) anticipates that biomarkers will become increasingly important in all phases of drug development [35]. A surrogate indicator is expected to predict the ultimate effect or safety of the therapy [36]. A range of CT methodological problems potentially interfere with validating biomarkers as indicators of the ultimate effectiveness or safety of a therapy [19, 24, 25].

OBJECTIVES

Given that cascading complexities from measurement error intrusions in CTs seemingly make changes to more accurate and precise outcome methods inevitable, we sought to understand what issues could emerge as CT investigators sought out alternative methodologies. Our aims are to encourage development of AD outcome methodologies able to avoid the cascade of complicating responses and compromised power of CTs, to avoid Type II errors as investigators interpret CT data analyses, and to bring increased scientific controls and systematizations into AD clinical research and patient care methods and practices.

METHODS

References were identified by searches of PubMed through March 2008 using "reliability", "accuracy", "precision", "Alzheimer’s disease clinical trials," “clinical trial methodology”, “biomarkers”, “surrogate endpoints”, and such terms. We used references in articles identified and those cited in earlier papers [3-5]. Only sources published in English were used. In preparations for this overview and commentary we placed special emphasis on methodologies needed to overcome limitations that remained unresponsive to or inadequately controlled with steps already proposed as improvements to current AD practices.[3]

RESULTS AND IMPLICATIONS

AD pathologies, especially neuronal death, impose irreversible losses on patients. Consequently, early diagnosis and, preferably, diagnoses reached prior to neuropathology associated with irreversible clinical changes deserve the highest research priorities. Dubois *et al*. [1] seek to diagnose AD and its pathologies prior to the point where irreversible change sets in. They emphasize biomarkers of AD as resources to aid early diagnoses and interventions and for their promise to reduce heterogeneity in study samples by selecting subjects based on quantitative deviations in specific biomarkers. Similarly, in our literature search, we found no methods other than biomarkers used as surrogate endpoints that could counter risks of Type II errors generated by flawed methodologies and practices.

The second pillar of effective drug development, providing maximum benefits in the clinic from available treatments, depends upon the clinician’s abilities to detect and quantify changes in pathological processes. This aim motivates authors to seek out criteria able to guide patient management [7, 16, 26, 30, 36] Given the inaccuracy and imprecision of clinical status ratings provided by trained researchers, it seems unreasonable to expect sufficient accuracy and precision from practitioners who must work without ongoing training and monitoring of their observational skills [7, 16]. Practitioners’ vulnerabilities to measurement errors and bias provide additional motivations to develop surrogate endpoints able to quantify drug effects on AD pathology and to introduce these endpoint measures into clinical practice. We foresee the need for AD drug developments that not only provide practitioners with efficacious drugs but also with the conditions of patient care needed to realize optimal effectiveness with drugs in patient care and the tools to realize these benefits.

Using these two contexts of drug development and clinical use, our reviews identified four core issues related to using surrogate markers. These factors are 1) limitations imposed by unavoidable human errors, 2) qualifying surrogate endpoints for AD research and patient care, 3) justifications for shifting to surrogate endpoints, and 4) effectively planning, implementing, pacing, and monitoring drug developments for quality.

The Limitations Imposed by the Ultimate Dependence of CTs on Rated Outcome Measures: Vulnerability to Human Error and Bias

The FDA and European authorities require, for new drug approvals, demonstrated quality of life benefits [35, 37, 38]. Faced with this ultimate dependence of all drug developments on clinically rated quality of life benefits, human error effects on ratings and inherent limitations in quality of life instruments will remain issues for drug developments. Thus imprecision, inaccuracies, and bias remain concerns for AD drug developments and clinical practice.

Averaging of values into means to analyze CTs reduces effects from imprecision and some forms of inaccuracy [16, 39]. Unfortunately, error sources produce much more consequential unreliabilities in assessments of individual patients where random error components are not routinely reduced by averaging as they are in CTs [5]. Of course, with excessive imprecision or inaccuracies, as demonstrated by Engelhardt *et al.* [7] and Cogger [8], the clinician cannot rationally guide patient management. Given the possibility raised by Demitrack [17] that some clinicians will not develop sufficient evaluation skills even with extensive training, we conclude that AD research must conservatively aim to provide clinicians with already adequately accurate and precise surrogate markers and clinical decision rules governing care decisions based on changes in markers [26]. In our review of the literature we found no other resources potentially able to provide practitioners with the gold standard accuracy and internally controlled precision needed to manage AD patients. The use of off-site expert raters being developed by Kobak *et al*. [9, 40] may prove useful for CTs but cannot provide a practical resource for clinicians.

Our review confirms the presence of currently unresolved problems of inaccuracy, imprecision, and bias potentially interfering with clinical ratings of AD patients, the consequent risks of Type II errors in CTs, and troubling implications for patient management. We propose to avoid these difficulties in clinical care by providing practitioners with surrogate markers linked to clinical decision rules. We did not uncover in the literature procedures and protocols for uses of surrogate markers adequate to insure clinicians' successes applying surrogate markers in patient care.

Qualifying Surrogate Endpoints for Applications in CTs and Patient Care

The AD community faces daunting tasks as it seeks surrogate endpoints: 1) the identification of biomarkers able to identify persons at high risk for AD; 2) the qualification of biomarkers as surrogate endpoints for preventive and therapeutic interventions; 3) standardization of methods for use of these surrogate markers in research and patient care, and 4) incorporation of well-evidenced methods and practices in protocols able to control practices in both research and patient care. These four aims alone do not guarantee clinical benefits and safety for AD drugs. Drugs potentially have activities, both associated with their principal mechanisms of action and inherent in the molecules, that carry deleterious effects. Temple [35] discusses numerous examples. In this section we assume the need to address issues and translations to clinical practice not always considered in discussions of surrogate endpoint developments. We propose a two pathway approach to development of surrogate endpoints. We base these proposals on our CT experiences and the sources in the literature.

As we discussed in Becker and Greig [3] problems from variance can be mitigated but not overcome using clinically skilled, well-trained, carefully monitored raters. We and Engelhardt *et al.* [7] see carefully prepared and executed CT designs as necessary to control variance and inaccuracy. Our response is to not leave clinical ratings of long-term benefits to less specialized researchers. We suggest reserving CT resources at specialized academic sites to evaluate long-term benefits crucial to evidencing the predictive powers of surrogate endpoints. In this way a pool of clinically experienced AD clinicians could be systematically trained and monitored for quality of rating performance. With experience accumulating over trials, AD research into surrogate endpoints would have available long-term benefit data collected under the best available conditions to control error interference. This effort seeks to maximize rater performance and to craft the research teams required to explore surrogate markers with confidence in the evaluation methodologies used.

In a second pathway CTs using already qualified surrogate endpoints could explore the efficacy of interventions. CTs using surrogate endpoints could be conducted in non-specialized research sites with increased confidence that failures are not Type II errors. Sponsors and investigators in these studies will have protocols for collection of samples or images, analyses of samples carried out in specialized laboratories, can be monitored using repeated sampling and spiked controls in each processing step, will have benefits of GLP to insure the accuracy and precision of data. Biomarkers with much reduced variances will provide the endpoints for comparison of active and placebo drug arms allowing smaller subject sample sizes without compromises to CT power. These CTs, with attention to dosing, drug interactions, effects from subsamples, and so forth can be designed and implemented to provide tests of drug efficacy, models for the conditions required in clinical practice to optimize this efficacy for each patient, and protocols and clinical decision rules to insure compliance in practice with the CT model. We suggest that these, equal, or more effective specializations of resources will be needed to address the realities of variance, bias, limited resources and so forth identified in the literature as potential risks to AD CT successes. Of course, even the concentration of best skilled clinical resources on the more difficult tasks of qualifying biomarkers as surrogate endpoints does not insure successes avoiding Type II errors.

Qualifying biomarkers as surrogate endpoints requires adequately reliable and precise evaluations of quality of life factors in individuals and evidence that the surrogate marker reliably predicts quality of life long-term benefits [35, 37, 41, 42]. Mere correlations between surrogate and clinical endpoints are not adequate [30, 43, 44]. In addition, not all statistically significant differences, even those for surrogate endpoints, will be inherently clinically important [44]. A statistical test of a difference between surrogate endpoint values in active and placebo drug treated groups provides only evidence of some difference without indicating clinical importance. One task accompanying development of protocols for using surrogate markers in AD will be to identify effect size estimators for markers to determine the importance of changes in the marker for the long-term outcomes of disease. Unfortunately, even with the interval or better data provided by biomarkers, quantitative changes in markers may not reflect parallel functional effects in brain systems. Thus the interval values available with many markers may express non-interval value associated consequences for brain functions and benefits. Effect size claims will need to be interpreted with these considerations taken into account.

Surrogate endpoints for efficacy do not avoid problems with drug safety. Problems with safety have been recently reported for drugs approved based on surrogate endpoints [34, 35]. Often these safety issues go undiscovered because of incomplete Phase III or lax Phase IV analyses. Academic AD medicine may have to challenge the elective status of Phase IV surveillance required by the FDA if AD medicine is to avoid similar safety problems after introduction of AD surrogate markers and approvals of drugs based on expectations associated with these markers.

Justifications for Using Surrogate Markers in AD Drug Regulatory Approvals

The FDA Modernization Act of 1997 allows for fast track approvals when a surrogate marker indicates the drug as most likely to safely provide clinical benefit for serious and life-threatening diseases [45]. Rationales under this Act are based on surrogate markers avoiding the delays as medicine awaits Phase III studies to demonstrate clinical benefits. We suggest the potential loss of AD drugs to Type II errors as an important additional practical rationale for pursuing early approvals of AD drugs because of the development of surrogate endpoints these approvals will require. We foresee clinically rated outcomes less dominant, biomarker research encouraged, the risk reduced for Type II errors, and hopefully new disease modifying AD drug approvals. Patients would not be well served by weakened regulatory standards, neither are they served well by drugs lost to Type II errors or underappreciated due to inadequately controlled CTs. We find in our earlier work and in the literature we reviewed for this paper no methods, other than the use of biomarkers, adequate to overcome variance effects in AD CTs.

One problem for AD investigators is evidence for more than one disease mechanism yielding AD neuropathology and clinical symptoms. Investigators must be open to biomarkers pointing towards subtypes within the current disorder [46]. If one pathological origin or process does not account for AD then a single biomarker will not monitor the disease process or capture the outcome until subdiagnostic types are identified [47]. It is possible that biomarkers may become surrogate endpoints for AD, for not yet identified AD subtypes, or not become surrogate substitutes for ratings of clinical benefits in AD yet come to play important roles advancing AD research, drug developments, and how practitioners manage therapies to most effectively intervene prior to a critical watershed, such as MCI impairments progressing to AD disabilities. AD researchers must not be limited to researching biomarkers solely as either surrogates for clinical outcomes or useful adjuncts to clinical assessment methods [25, 48]. We may not know the utility of biomarkers in AD until we have well-characterized and researched biomarkers to work with in research and patient care. Issues of surrogate endpoint qualification await successes in biomarker developments such as ADNI investigators presently seek.

Monitoring Researchers and Sites for Quality Assurance

As we and others have already discussed and documented, one of the most important aspects in planning, implementing, pacing, and monitoring drug developments for quality involves the more effective training and monitoring of investigators at sites for quality assurance. We found in our review of the literature no reasons to reduce our concerns that important risks to validity of CTs go unaddressed by sponsors and investigators. Our earlier work and the background to this paper document the complexities that must be considered if investigators are to be freed from concerns that drug failures in development are possible due to methods and practices. We remain impressed with the abilities of Engelhardt *et al.* [7] to monitor closely raters at distant sites for compliance with protocols. From our own experiences with trials and participation in trials sponsored by others we can witness that, over decades, evaluations of adequacy of training in rating instruments have used reliabilities, statistics that at best are insensitive to imprecision and inaccuracies shown to adversely affect CT power. Monitoring has been too narrowly focused on detecting falsified data without attention to eradicating practices that increase error intrusions. As a consequence, we foresee the need for more detailed protocols governing all procedures that could put a CT at risk and for training and monitoring at the intensities needed to insure the validity of data sets.

CONCLUSIONS

The human errors interfering in current AD CTs indicate two action strategies for AD clinical pharmacologists. Because qualifying surrogate markers requires sound data about long-term benefits, the AD clinical pharmacology community needs to organize resources to provide this required data. We suggest commitment of the most clinically experienced and clinical rating well-trained investigators to studies to demonstrate the predictive powers of candidate surrogate endpoints.

We expect that qualification of biomarker candidates as surrogate AD endpoints will potentially overcome current Type II error risks to CTs. This advance depends on the conduct of CTs being adequately governed by appropriate protocols, training of investigators, and monitoring of sites for compliance with all study protocols. Since biomarkers can reduce cascading risks from variance, we see CT efficacy testing safely carried out at less specialized sites.

Divisions of labor are not novel; how we propose to use divided resources is less usual. Cancer researchers accept, for drugs granted accelerated approval based on demonstrated effectiveness against a surrogate marker, two-stage CT progressions towards final regulatory approvals [49]. Final regulatory approval awaits longer-term follow-ups of patients in CTs and systematic post-marketing surveillance to confirm the power of the surrogate marker to predict patient benefits and drug safety. We are proposing a somewhat more specialized progression for AD. In current practice a CT may fail due to the drug not reaching target sites at optimal concentrations, the target not being linked to disease expression or progression, or failed methodologies. In the division of labor model we propose drug effects on targets are confirmed by biomarker changes and the links of these drug targets to clinical status demonstrated using AD research's most skilled investigators. For subsequent candidate compounds the biomarker becomes the indicator for drug effects at the molecular target. Phase III studies may fail to demonstrate surrogate endpoint changes or in followups fail to demonstrate clinical disease effects. In each instance, the presence of the established theory linking the surrogate endpoint to molecular targets and disease effects shifts attention first to possible methodological failures, including undetected compound linked factors. An alternative explanation, that surrogate endpoint changes are not part of the disease process, undermines the theory of disease. We suggest this division of labor model as one way to reintroduce into AD clinical pharmacology some systematic uses of therapeutics as a means to understanding disease mechanisms.

The concentration of specialized resources at experienced academic sites aims to minimize risks due to inaccuracies, imprecision, and biases interfering with demonstrated long-term predictive power of potential surrogate endpoints [50]. Commercial firms tend not to validate surrogate endpoints since this only eases developmental tasks for competitors. We earlier suggested that all research sites, investigators, and raters, those who will become involved in surrogate marker qualification studies and those that will carry out efficacy CTs, participate in each of the Phases I, II, and III of a drug's development to provide cumulative experience and training in uses of instruments and tests to be employed in later CTs. Based on performance, drug development researchers can drop from further participation raters and sites that do not reach required levels of reliability.

We emphasize one further note of caution based on our work and the literature. Typically, rating scale reliabilities may be high; yet surprisingly, not reflect imprecision and inaccuracies that encourage Type II errors in CTs and undermine an individual's care in clinical practice [7, 16, 51, 52]. More focused attention needs to be given to eliminating the inaccuracies, imprecision, and biases that risk Type II errors in spite of misplaced confidence in high reliabilities and compliance with randomization in CTs [16, 53]. Given the FDA's and medicine's interest in assuring that treatments provide clinical benefits and safety for patients, quality of life ratings will always provide the ultimate tests of superiority for many medical interventions. Accuracy, precision, freedom from bias, and other factors will continue to limit the utilities of these measures.

Researchers need to expand the horizons of their thinking to go beyond the CT as providing efficacy evidence. CTs offer comprehensive opportunities for both witnessing to efficacy and providing the practices that will optimize the efficacy of a drug for the patient in clinical practice. Hopefully, parallel to the innovations we encourage for demonstrating efficacy, better systematized Phase IV monitoring will differentiate drugs that fail due to safety lapses from lapses of protocols that allow practices that cause drugs to become unsafe or ineffective. A rational AD clinical pharmacology would in our view ask for more rigorous research practices and greater regulatory controls to assure three changes: first that drugs are tested such that methodologies can not undermine efficacy, second that drugs are tested using biomarker surrogate endpoints as drug targets, and lastly that drugs are used by practitioners in the contexts, conditions, and with surrogate endpoints with which the drugs were proven optimally safe and effective.

ACKNOWLEDGMENTS

This work was supported in part by the Intramural Research Program of the National Institute on Aging, NIH.

Contributors

Each of the authors contributed to the conception and writing of the paper.

ABBREVIATIONS

AD	=	Alzheimer's disease
ADNI	=	Alzheimer's disease Neuroimaging Initiative
ChEI	=	Cholinesterase inhibitor

CT	=	Clinical trial
FDA	=	Food and Drug Administration
GLP	=	Good laboratory practices
HDS	=	Hamilton Depression Scale
MRI	=	Magnetic resonance imaging
MCI	=	Mild Cognitive Impairment
PET	=	Positron emission tomography
SPECT	=	Single positron emission computerized tomography

REFERENCES

[1] Dubois, B.; Feldman, H.H.; Jacova, C.; DeKosky, S.T.; Barberger-Gateau, P.; Cummings, J.; Delacourte, A.; Galasko, D.; Gauthier, S.; Jicha, G.; Meguro, K.; O'Brien, J.; Pasquier, F.; Pobert, P.; Rossor, M.; Salloway, S.; Stern, Y.; Visser, P.J.; Scheltens, P. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.,* **2007**, *6*, 734-46.

[2] Pangalos, M.N.; Schechter, L.E.; Hurko, O. Drug development for CNS disorders: strategies for balancing risk and reducing attrition. *Nat. Rev. Drug Discov.,* **2007**, *6*, 521-32.

[3] Becker, R.E.; Greig, N.H. Alzheimer's disease drug development in 2008 and beyond: problems and opportunities. *Curr. Alz. Res*., **2008**, *5*, 346-57.

[4] Becker, R.E.; Greig, N.H.; Giacobini, E. Why do so many drugs for Alzheimer's disease fail in development? Time for new methods and new practices? *J. Alzheimers Dis*., **2008**, *15*, 303-25.

[5] Becker, R.E. Lessons from Darwin: 21st century designs for clinical trials. *Curr. Alz. Res*., **2007**, *4*, 458-67.

[6] Chalmers, I. Underreporting research is scientific misconduct. *JAMA*, **1990**, *263*, 1405-8.

[7] Engelhardt, N.; Feiger, A.D.; Cogger, K.O.; Sikich, D.; DeBrota, D.J.; Lipsitz, J.D.; Kobak, K.A.; Evans, K.R.; Potter, W.Z. 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*, 71-4.

[8] Cogger, K.O. Rating rater improvement: a method for estimating increased effect size and reduction of clinical trial costs. *J. Clin. Psychopharmacol*., **2007**, *27*, 418-20.

[9] Kobak, K.A.; DeBrota, D.J.; Engelhardt, N.; Williams, J.B.W. 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**, June, Abstract.

[10] Williams, J.B.W. A structured interview guide for the Hamilton Depression Rating Scale. *Arch. Gen. Psychiatry,* **1988**, *45*, 742-7.

[11] Weiner, M.W. Alzheimer's Disease Neuroimaging Initiative. *Int. Conf. Alzheimer's Disease. Chicago IL.*, **2008**, July, Abstract S1-01-01.

[12] Weiner, M. Appendix. Introductory section background and significance of the overall proposal, **2004**, Hppt:www.loni.ucla.edu/ADNI/pdf/ADNI_Introduction.pdf. Viewed Nov. 28, 2007.

[13] Raschetti, R.; Albanese, E.; Vanacore, N.; Maggini, M. Cholinesterase inhibitors in mild cognitive impairment; a systematic review of randomized trials. *PLoS Med*., **2007**, *4*, e338.doi:10.1371/journal.pmed.0040338.

[14] Visser, P.J.; Scheltens, P.; Verhey, F.R. Do MCI criteria in drug trials accurately identify subjects with predementia Alzheimer's disease? *J. Neurol. Neurosurg. Psychiatry,* **2005**, *76*, 1348-54.

[15] Foster, N. Imaging advances for early detection and new drug assessment in Alzheimer's disease. *Presentation at Clinical Neurosciences Center University of Utah.,* **Nov**. **28**, **2007**.

[16] Becker, R.E.; Markwell, S. Problems arising from generalizing of treatment efficacy from clinical trials in Alzheimer's disease. *Clin. Drug Invest*., **2000**, *19*, 33-41.

[17] Demitrack, M.A.; Faries, D.; Herrera, J.M.; DeBrota, D.; Potter, W.Z. The problem of measurement error in multisite clinical trials. *Psychopharmacol. Bull*., **1998**, *34*, 19-24.

[18] Targum, S.D. Evaluating rater competency for CNS clinical trials. *J. Clin. Psychopharmacol*., **2006**, *26*, 308-10.

[19] Butcher, J. The hunt for drugs to modify Alzheimer's disease. *Lancet Neurol*., **2007**, *6*, 1038-9.

[20] Jansen, R.L.; Teeter, J.G.; England, R.D.; White, H.J.; Pickering, E.H.; Crapo, R.O. Instrument accuracy and reproducibility in measurements of pulmonary function. *Chest*, **2007**, *132*, 367-8.

[21] Colburn, W.A. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. *J. Clin. Pharmacol*., **2000**, *40*, 1419-27.

[22] Code of Federal Regulations. Good laboratory practice standards. *Code of Federal Regulations*, **1997**, Title 40, Part 792, Chapter I, Sections 792.1- 792.195.

[23] Nordic Council on Medicines. *Good clinical trail practice: Nordic Guidelines. NLN Publication No. 28*, Uppsala, Sweden, **1989**.

[24] Thal, L.J.; Kantarci, K.; Reiman, E.M.; Klunk, W.E.; Weiner, M.W.; Zetterberg, H.; Galasko, D.; Pratico, D.; Griffin, S.; Schenk, D.; Siemers, E. The role of biomarkers in clinical trials for Alzheimer disease. *Alzheimer's Dis. Assoc. Disord.*, **2006**, *6*, 6-15.

[25] Vellas, B.; Andrieu, S.; Sampaio, C.; Coley, N.; Wilcock, G.; European Task Force Group. Endpoints for trials in Alzheimer's disease: a European task force consensus. *Lancet Neurol*., **2008**, *7*, 436-50.

[26] Becker, R.E. Focusing pharmaceutical research on patient care: using clinical trials to develop clinical decision rules. *Clin. Drug Invest*., **2002**, *22*, 269-78.

[27] Nordberg, A. PET imaging of amyloid in Alzheimer's disease. *Lancet Neurol*., **2004**, *3*, 519527.

[28] Grundman, M.; Jack, C.R.; Petersen, R.C.; Kim, H.T.; Taylor, C.; Datvian, M.; Weiner, M.F.; DeCarli, C.; DeKosky, S.T.; van Dyck, C.; Darvesh, S.; Yaffe, K.; Kaye, J.; Ferris, S.H.; Thomas, R.G.; Thal, L.J.; The Alzheimer's Disease Cooperative Study. Hippocampal volume is associated with memory but not non-memory cognitive performance in patients with mild cognitive impairment. *J. Mol. Neurosci*., **2003**, *20*, 241-8.

[29] Feldman, H.H.; Ferris, S.; Winblad, B.; Sfikas, N.; Mancione, L.; He, Y.; Tekin, S.; Burns, A.; Cummings, J.; del Ser, T.; Ogogozo, J-M.; Sauer, H.; Scheltens, P.; Scarpini, E.; Herrmann, N.; Farlow, M.; Potkin, S.; Charles, H.C.; Fox, N.C.; Lane, R. Effect of rivastigmine on delay to diagnosis of Alzheimer's disease from mild cognitive impairment: the InDDEx study. *Lancet Neurol*., **2007**, *6*. 501-12.

[30] Frank, R.; Hargreaves, R. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov*., **2003**, *2*, 566-80.

[31] Jansen, R.L.; Teeter, J.G.; England, R.D.; Howell, H.M.; White, H.J.; Pickering, E.H.; Crapo, R.O. Sources of long-term variability in measurements of lung function: implications for interpretation and clinical trial design. *Chest*, **2007**, *132*, 396-402.

[32] Jones, D.W.; Appel, L.J.; Sheps, S.G.; Roccella, E.J.; Lenfant, C. Measuring blood pressure accurately: new and persistent challenges. *JAMA*, **2003**, *289*, 1027-30.

[33] Lappin, G.; Kuhnz, W.; Jochemsen, R.; Kneer, J.; Chaudhary, A.; Oosterhuis, B.; Drijfhout, W.J.; Rowland, M.; Garner, R.C. Use of microdosing to predict pharmacokinetics at the therapeutic dose: Experience with 5 drugs. *Clin. Pharmacol. Ther*., **2006**, *80*, 203–15.

[34] Lesko, L.J.; Atkinson, A.J. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Ann. Rev. Pharmacol. Toxicol.*, **2001**, *41*, 347-66.

[35] Temple, R. Are Surrogate Markers Adequate to Assess Cardiovascular Disease Drugs? *JAMA*, **1999**, *282*, 790-5.

[36] Qaseem, A.; Snow, V.; Cross, J.T.; Forciea, M.A.; Hopkins, R.; Shekell, P.; Adelman, A.; Mehr, D.; Schellhase, K.; Campos-Outcait, D.; Santaguida, P.; Owens, D.K.; and the Joint American College of Physicians/American Academy of Family Physicians Panel on Dementia. Current pharmacological treatment of dementia: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. *Ann. Intern. Med*., **2008**, *148*, 370-8.

[37] Katz, R. FDA evidentiary standards for drug development and approval. *NeuroRx*., **2004**, *1*, 307-16.

[38] Leplege, A.; Hunt, S. The problem of quality of life in medicine. *JAMA*, **1997**, *278*, 47-50.

[39] Perkins, D.O.; Wyatt, R.J.; Bartko, J.J. Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials. *Biol. Psychiatry*, **2000**, *47*, 762-6.

[40] Kobak, K.A.; Engelhardt, N.; Lipsitz, J.D. Enriched rater training using Internet-based technologies: a comparison to traditional rater training. *J. Psychiatr. Res*., **2006**, *3*, 192-9.

[41] D'Agostino, R.B. Debate: the slippery slope of surrogate outcomes. *Curr. Control Trials Cardiovasc. Med*., **2000**, *1*, 76-8.

[42] Prentice, R.L. Surrogate endpoint in clinical trials: definition and operational criteria. *Stat. Med*., **1989**, *8*, 431-40.

[43] Baker, S.G.; Kramer, B.S. A perfect correlate does not a surrogate make. *BMC Med. Res. Methodol.*, **2003**, *3*, 16. http://www.biomed central.com/1471-2288/3/16 Viewed March 10, 2008.

[44] Pincus, T.; Stein, C.M. Why randomized controlled clinical trials do not depict accurately long-term outcomes in rheumatoid arthritis: some explanations and suggestions for future studies. *Clin. Exp. Rheumatol*., **1997**, *15 Suppl. 17*, S27-S38.

[45] Food and Drug Administration. *Modernization Act of 1997*. http://www.fda.gov/opacom/7modact.html. Accessed March 10, **2008**.

[46] Aronson, J.K. Biomarkers and surrogate endpoints. *Brit. J. Clin. Pharmacol.*, **2005**, *59*, 491-4.

[47] Bucher, H.C.; Guyatt, G.H.; Cook, D.J.; Holbrook, A.; McAlister, F.A.; for the evidence-based medicine working group. User's guides to the medical literature XIX. Applying clinical trial results. *JAMA*, **1999**, *282*, 771-8.

[48] Galasko, D. Biological markers and the treatment of Alzheimer's disease. *J. Mol. Neurosci*., **2001**, *17*, 119-125.

[49] Kelloff, G.J.; Bast, R.C.; Coffey, D.S.; D'Amico, A.V.; Kerbel, R.S.; Park, J.W.; Ruddon, R.W.; Rustin, G.J.S.; Schilsky, R.L.; Sigman, C.C,; Vande Woude, G.F. Biomarkers, surrogate endpoints, and

the acceleration of drug development for cancer prevention and treatment: an update. *Clin. Cancer Res*., **2004**, *10*, 3881-4.

[50] Lonn, E. The use of surrogate endpoints in clinical trials: focus on clinical trials in cardiovascular diseases. *Pharmacoepidemiol. Drug Saf*., **2001**, *10*, 497-508.

[51] El Achhab, Y.; Nejjari, C.; Chikri, M.; Lyoussi, B. Disease-specific health-related quality of life instruments among adults diabetic: a systematic review. *Diabetes Res. Clin. Pract.,* 2008, *80*, 171-84.

[52] Schrag, A.; Selai, C.; Mathias, C.; Low, P.; Hobart, J.; Brady, N.; Quinn, N.P. Measuring health-related quality of life in MSA: the MSA-QoL. *Mov. Disord*., **2007**, *22*, 2332-8.

[53] Worrall, J. Why there's no cause to randomize. *Brit. J. Phil. Sci*., **2007**, *58*, 451-88.

Received: December 15, 2008 **Revised: December 24, 2008** **Accepted: December 24, 2008**