FDA: Evidentiary Standards for Drug Development and Approval

Russell Katz

Division of Neuropharmacological Drug Products, United States Food and Drug Administration, Rockville, Maryland 20852

Summary: The United States Food and Drug Administration is charged with approving drug treatments that have been shown to be safe and effective. Relevant statutes and regulations provide a legal framework for establishing safety and effectiveness that is sufficiently flexible to ensure that appropriate scientific data are collected for specific treatments targeted to particular diseases. Nonetheless, all clinical trials proposed to establish effectiveness must incorporate common elements in order for the appropriate legal and scientific

standards of drug approval to be met. This article will discuss the relevant laws and regulations pertaining to the current effectiveness standard and will discuss the most important clinical trial design elements currently considered acceptable for applications for treatments of neurologic and psychiatric illness. **Key Words:** Substantial evidence of effectiveness, statistical analyses, adequate and well-controlled clinical investigations, active control trials, disease progression.

In 1962, the Federal Food, Drug, and Cosmetic Act (the Act), the primary statutory basis for drug approval in the United States, was amended under the so-called Kefauver-Harris amendments. These amendments mandate that the Food and Drug Administration (FDA; the Agency) must determine that a drug product is both safe and effective before it may be approved for marketing. Before 1962, the Act required only that a drug product be safe for use before it could be approved for marketing. Importantly, in addition, these amendments also gave the FDA authority to regulate research in humans with investigational drugs; specifically, manufacturers could no longer move drugs in interstate commerce without full market approval or, for an unapproved drug, permission from the FDA in the form of an exemption from this prohibition, a process that eventually evolved into the Investigational New Drug (IND) process. In this paper, I will focus on the evidentiary standard for drug approval in the United States, with particular emphasis on the effectiveness standards and how they apply to drugs to treat neurologic and psychiatric illness.

As noted above, a drug product must be found to be effective and safe before it may be approved for general marketing. The Act explicitly gives the legal definition of the evidence necessary for the Agency to determine that a drug product has been found to be effective; that standard is “substantial evidence of effectiveness,” and is defined as “evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof.”

“Adequate and well-controlled” investigations are further defined in Section 314.126 of Title 21 of the Code of Federal Regulations (CFR) (21 CFR 314.126). The CFR is a compendium of regulations promulgated by the Agency to carry out the provisions of the Act. Regulations are adopted by federal agencies through a process of notice and comment rule-making, as opposed to statutes, which become law through an Act of Congress. In any event, regulations have the force of law.

Here, the essential elements of trial designs that the Agency accepts as fulfilling the requirement of the statute for adequate and well-controlled investigations are explic-

The views expressed in this article are those of the author, and are not an official statement of the Food and Drug Administration.

Address correspondence and reprint requests to Russell Katz, M.D., Division of Neuropharmacological Drug Products, U.S. Food and Drug Administration, Center for Drug Evaluation and Research, 1451 Rockville Pike, Room 4037, Rockville, MD 20852. E-mail: katzr@cder.fda.gov.

itly described. These essential elements include the following: 1) a clear statement of the objective of the study and a summary of the methods of analysis of the trial results, 2) the design must permit a valid comparison with a control group to permit a quantitative assessment of the effect of the drug, and 3) the protocol for the study should precisely define the design, including the duration of the study, whether treatments are parallel or sequential, and issues related to sample size.

Critically, the regulations define the five types of control groups that may be considered acceptable. The following first four controls are concurrent (they are included in the same study):

- 1) placebo concurrent control; patients are assigned to treatment or placebo (usually through randomization),

- 2) dose comparison concurrent control; patients are assigned to one of several dose levels of the treatment,

- 3) no treatment concurrent control; patients are assigned to treatment or no treatment,

- 4) active treatment concurrent control; patients are assigned to proposed treatment or other active treatment,

- 5) historical control; patients are assigned to proposed treatment only (their responses are compared to those of similar patients not in the trial).

Any given trial may incorporate several of these controls (for example, a given study may include placebo and an active treatment control).

Additional design elements required by the regulations include a description of the method of selection of patients (to ensure that they have the condition under study), a description of the method of assigning treatments (e.g., randomization), a description of the methods used to minimize bias on the part of patients, observers, and data analysts (e.g., blinding), and a description of the assessment of the patients' responses. We will return later to a more detailed discussion of some of these requirements.

The use of the word "clinical" in the definition of substantial evidence has routinely been interpreted as "human." That is, a drug may not be approved unless there is evidence from studies in humans that the drug has the claimed effect. The word "investigations" has also been interpreted routinely as requiring more than one such investigation. That is, this specific word has served as the basis for the standard Agency requirement for independent replication or corroboration of a finding seen in a single study before it may be approved (strict replication, i.e., the repeating of trials with identical design, can be acceptable, but data from independent trials of different design is preferred, not only because this reduces the possibility that an unknown bias in a single trial will be repeated but also because different information can be obtained from different trials, e.g., different disease severities can be studied). While these criteria (the requirement for more than one study in humans) are

still very much the standard for drug approval, recent changes to the Act and regulations have introduced the possibility that, in rare cases, different standards may apply.

In 1997, the Act was amended via passage of the Food and Drug Administration Modernization Act (FDAMA). The FDAMA introduced a number of important changes into the Act, but for our purposes here the important revision included a revised definition of substantial evidence of effectiveness.

Under FDAMA, the definition of substantial evidence of effectiveness was amended to include a single adequate and well-controlled study and "confirmatory evidence," if the Agency determines, "based on relevant science," that the data establish effectiveness. While Congress gave no guidance about when such a standard should be applied, or what could constitute confirmatory evidence, the Agency has produced a document (Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products)¹ that provides a detailed discussion of the regulatory and scientific considerations that are involved in the decision to approve an application on the basis of a single adequate and well-controlled clinical trial.

This document describes two categories of trials in which a single study might serve as substantial evidence of effectiveness; those cases in which a single study may receive substantiation from related data, and those cases in which there is no independent data outside the trial that can provide substantiation. In the former category, examples would include studies of new dosage regimens or forms (e.g., controlled-release products when a drug is only approved as an immediate release formulation) for treatments already approved, studies in related populations (e.g., studies in pediatric patients when the drug is approved in adults), studies performed under related conditions of use (e.g., use in monotherapy when the drug is approved as adjunctive therapy), and studies in different severity strata (e.g., studies in severely ill patients in which the drug is approved only in mildly ill patients).

In the second category, examples might include large multicenter studies in which several individual centers show independent evidence of effectiveness or such studies that provide evidence of effectiveness on multiple "independent" outcomes (e.g., a clinical and an imaging outcome). It is critical to understand that this standard is considered to apply in only the rarest settings (perhaps for a treatment in which a single trial would be unethical to repeat) and, to date, has not been applied to treatments for neurologic or psychiatric disease.

Recently, the Agency has adopted a regulation under 21 CFR 314.600 (Subpart I, Approval of New Drugs When Human Efficacy Studies Are Not Ethical or Feasible, the so-called "Animal Rule") that suspends the requirement for evidence of effectiveness to be generated

in humans. This rule applies to treatments designed to prevent or ameliorate serious or life-threatening conditions caused by exposure to lethal or disabling "...toxic, biological, chemical, radiological, or nuclear substances." This rule applies only in those cases in which it would be unethical to deliberately expose humans to these agents and field studies to evaluate the effectiveness of the treatment "...have not been feasible." In these cases, studies performed in animals can serve as the basis of approval, if the Agency determines that these data are "...reasonably likely to produce clinical benefit in humans." The rule applies if there is a reasonably well understood mechanism of the toxicity of the substance and its prevention or treatment by the drug, the beneficial effect of the drug is demonstrated in multiple animal species, or in a species that is most clearly relevant to humans, the endpoint in the animal studies is related to the desired human endpoint (generally mortality or major morbidity), and the data permit the selection of a dose in humans. To date, the Agency has approved a single application on the basis of this rule, for the use of pyridostigmine bromide as a pretreatment for nerve agent poisoning.

Other statutory and regulatory provisions are relevant to a discussion about evidentiary requirements for the approval of drugs to treat neurologic and psychiatric disease.

Perhaps most important among these are the provisions related to surrogate endpoints. There is considerable interest currently in approval of treatments on the basis of their effects on surrogate endpoints for several reasons, including the view that studies that rely on surrogate measures as their primary outcome measures can be smaller (fewer patients) and shorter than studies that rely on an effect on a more traditional outcome. More important, perhaps, is the view that an effect on an appropriately chosen surrogate marker may be reflective of an effect on the underlying pathology of a disease and not just a symptomatic effect.

Surrogate endpoints can be defined as laboratory measures or other tests that have no direct or obvious relationship to how a patient feels or to any clinical symptom, but on which a beneficial effect of a drug is presumed to predict a desired beneficial effect on such a clinical outcome. For our purposes, two "types" of surrogate outcomes can be defined: "validated" surrogate outcomes, and "unvalidated" surrogate outcomes.

Validated surrogate outcomes are those tests for which there is adequate evidence that a drug effect on the measure predicts the clinical benefit desired. The Agency has approved many treatments on the basis of their effects on validated surrogate measures. For example, antihypertensives are approved on the basis of their effects on a measurement (blood pressure), not on any effect on a symptom that is detectable to a patient (typically, ele-

vated blood pressure is asymptomatic). These drugs are approved, however, because evidence has demonstrated that lowering blood pressure in the long run has beneficial clinical effects (on, for example, the risk for stroke and myocardial infarction). Similarly, treatments for elevated cholesterol are approved on the basis of their effects on validated surrogates, as are other treatments.

Unvalidated surrogates, on the other hand, are measures for which evidence does not exist that a drug effect on the measure predicts the desired clinical outcome.

In the wake of the HIV/acquired immune deficiency syndrome (AIDS) tragedy, however, in 1992 the Agency adopted a new regulation (21 CFR 314.500, Subpart H, Accelerated Approval of New Drugs for Serious or Life-Threatening Illnesses), in which, for the first time, approval of a treatment on the basis of its effects on an unvalidated surrogate was permitted.

The regulation permits approval of a drug on the basis of clinical trials (which must be adequate and well-controlled) in serious or life-threatening illnesses that establish that "...the drug product has an effect on a surrogate endpoint that is reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence, to predict clinical benefit...". This regulation applies only to treatments that offer a meaningful therapeutic benefit over that provided by available products (...e.g., ability to treat patients unresponsive to, or intolerant of, available therapy, or improved patient response over available therapy)." The regulations require that the surrogate marker be validated in studies completed after marketing, and, if this is not accomplished, the Agency may remove the product from the market in an expedited manner. The specific provisions (that the regulation be applied only for serious illnesses and for treatments with a major impact) reflect the Agency's acknowledgment that these approvals introduce a level of uncertainty into the approval process that is ordinarily not present (namely, the uncertainty that the effect of the drug on the surrogate will predict the desired clinical benefit), and that this level of uncertainty is acceptable only in the context of truly important treatments. As part of FDAMA, similar language permitting drug approval on the basis of an effect on such surrogates was added to the Act.

The consideration of uncertainty in drug approval provides a convenient window through which to glimpse important general principles that the Agency routinely applies in drug approval.

Of course, no scientific conclusion can be considered certain. Any conclusion reached on the basis of evidence generated in well designed and conducted experiments (clinical investigations) can be wrong in any given case. Indeed, there are four explanations for any difference seen between treatment and control groups: fraud, bias, chance, and drug effect. The Agency routinely seeks to

minimize the likelihood that any beneficial effect seen in a drug trial is the result of the first three possibilities.

Although fraud in the conduct and reporting of clinical trials is rare, it does occur. For those studies in new drug applications that are considered to provide critical effectiveness data, the Agency inspects the records of selected sites (these may be sites with the largest number of patients, sites with the largest treatment effects, or sites chosen based on other considerations), to ensure that the data were accurately recorded and reported. It must be acknowledged, however, that resource limitations preclude detailed inspections of most of the clinical data generated, and the Agency relies on frequent monitoring of study sites by manufacturers and local institutional practices to minimize the possibility of fraud and to detect it in those rare instances in which it occurs (these various inspectional activities also are expected to detect errors in data recording and reporting, which happen frequently, and which ordinarily have no material effect on the conclusions reached).

Great effort is made to minimize bias in the conduct of clinical trials and the analysis of clinical trial data. Most of the Agency's efforts in this regard are directed at the review of the protocol, with special attention to the various elements of trial design that may have an impact on this question (with particular emphasis on the efforts made by the sponsor to protect the blind). There are occasions in which maintaining double-blind conditions are not considered feasible; in these cases, it is usually imperative that the critical outcome data are reviewed and assessed by outside assessors who are blind to treatment assignment. A recent case illustrates this point. A comparative trial was conducted comparing the effects of clozapine (an atypical antipsychotic drug) to another atypical antipsychotic drug on reducing suicidal ideation and behavior. Because treatment with clozapine requires frequent (every week to every 2 weeks) blood tests to monitor for agranulocytosis, it was felt to be unethical to perform these frequent tests on the patients randomized to the comparator drug; hence, this effectively made the study unblinded. However, the data on the outcomes (descriptions of episodes of suicidal behavior) were submitted to a panel of outside expert assessors who were blinded to treatment assignment. Other trials in which the outcomes are considered "hard" and objective outcomes (for example, mortality) may be performed in an unblinded manner, if ethical (or overwhelming practical) considerations make blinding impossible, but only if it is decided that unblinding cannot effect the outcome (one can easily imagine that, in a particular case, knowledge of treatment assignment, either on the part of the patient or the assessor, could have an important effect on the outcome, even if the outcome is mortality or another "objectively" measured outcome).

Trials of neurologic and psychiatric treatments essen-

tially always employ randomized assignment to study treatment (not only does randomization ensure, as much as possible, an equitable distribution of relevant characteristics among the treatment groups, but it also provides the theoretical support for the use of most statistical analytic techniques). Further, these trials invariably employ parallel treatment groups (in which each patient is assigned to only one study treatment). Although there is no prohibition against the use of cross-over studies (in which patients are assigned to receive all treatments, in a random order), their interpretation is complicated (due, for example, to progression of the disease during the trial and potential residual effects of a treatment into subsequent treatment periods), and for this reason they are not generally relied upon. In almost all cases, the Agency relies on the results of analyses of a so-called modified intent-to-treat (ITT) population. That is, all patients who receive at least one dose of study drug and receive at least one outcome assessment are included in the efficacy analysis. While analyses of this population are not free of problems (especially if the number of dropouts is large), they have the great advantage of including the groups most representative of the randomized groups. Analyses that include only a subset of the entire randomized group (for example, only those patients who complete the trial) have the great disadvantage of introducing a potential bias into the results, because these groups are no longer fully randomized. For this reason, while analyses of the ITT population may be flawed in a given case, these analyses are usually considered primary.

It is always possible that a beneficial outcome could have occurred as a chance finding. Detailed statistical plans and appropriate analytic methodologies are applied to the analysis of any trial so that the level of chance that could account for the results is minimized.

It is widely known that the Agency will conclude that a trial is considered "positive" if the p value generated for the between-treatment contrast is less than or equal to 0.05. What does this mean?

First, it is important to note that the statistical analysis of any trial starts with the articulation of the "null hypothesis," a statement that the treatment and the control group are not different. After an appropriate analytic method is applied to the data, a p value is generated that describes the quantitative comparison between the treatment groups. The p value is the probability that the difference seen (or one more extreme) was due to chance, if the null hypothesis was true. That is, under the hypothesis that the treatments are not different, the p value describes how likely the difference seen could have occurred by chance. Obviously, the smaller the probability (the smaller the p value), the more likely the treatments are not the same (the null hypothesis is "rejected") and therefore, the more likely we are correct in assuming the drug is "effective." The probability that we

will falsely conclude that the drug is effective when in fact it is not (that is, if the null hypothesis is true) is called the type I error, and, as we have been discussing, the typical “cap” on the type I error rate is set at 5% (a p value of 0.05). A between-treatment difference that yields a p value of 0.05 or smaller is referred to as statistically significant.

It is important to recognize that no law or regulation requires this (or any other specific) statistical standard to be applied to the analysis of clinical trials (although we have seen that the regulations require a quantitative comparison of the effects of the drug to those of a control group). The adoption (and general application) of this standard has evolved as a reflection of the scientific community’s view that this degree of possibility that we could falsely conclude that the drug is effective when in fact it is not is acceptable. Certainly, there are cases in which the type I error is permitted to be (somewhat) greater than 5% (depending upon numerous conditions, this may be acceptable), but, from a regulatory perspective, it is critical that a standard be in place (one can easily imagine the regulatory chaos that would ensue if no generally accepted standard for deciding when a study was positive existed). The usual statutory requirement for independent replication or corroboration further ensures that a finding that is statistically significant is not a chance finding.

Considerable effort is expended in the drug review process (both at the protocol as well as at the data review level), in ensuring that the true type I error rate in a clinical trial is held at or less than 5%, i.e., the type I error rate is not “inflated.” There are numerous common analytic and design elements of clinical trials that can result in the generation of a p value of 0.05 or less, but that in fact disguise the fact that the type I error rate for the trial as a whole is considerably higher than 5% (these p values that appear to be 0.05 or less but that in fact are misleading are called “nominally significant”; i.e., they are 0.05 or less in name only, not in reality).

Typically, the Agency requires that, for definitive effectiveness trials, protocols prospectively specify a single primary outcome, i.e., a single outcome is designated as the outcome on which the primary between-treatment contrast is performed, and the results of this analysis serve as the basis for concluding that the study can contribute to a finding of substantial evidence of effectiveness (a so-called “pivotal” study). There are examples, however, of situations in which the Agency requires that a statistically significant difference between treatments on two outcomes be shown. The most common example is in studies of treatments with patients with Alzheimer’s disease, in which the typical study has two primary outcome measures: a measure of cognitive functioning (to assess the “core” symptoms of the disease) and a measure of global functioning (to ensure that the

change in cognitive functioning produces a clinically meaningful effect). In this case, a statistically significant difference between treatments must be obtained on analyses of both outcome measures. When multiple outcome measures are designated as “primary” by sponsors, when multiple time points are designated as important times to assess the drug-control differences, when multiple doses of the active treatment are compared to the control, or when multiple analytic methods are proposed, the opportunity arises for the generation of multiple “nominally” significant p values. All of these scenarios increase the likelihood that the type I error will be inflated and, therefore, increase the likelihood that we will falsely conclude that the treatment is effective. Because this is an outcome to be assiduously avoided, great care and effort are expended to ensure that protocols are explicitly constructed to include plans for the statistical analysis of trials that will “preserve” the type I error rate.

The statistical approach to analyzing clinical trials used by the Agency (almost universally), is based on what is known as “frequentist” statistics. This paradigm utilizes probabilities to assess the possibility that the results seen were a chance finding (the type I error rate we have been discussing), as well as to assess the strength of that finding (the smaller the p value, the greater the strength of the evidence). These techniques are not without theoretical problems, and there are other statistical paradigms that exist and that have been proposed to replace frequentist approaches. The most commonly discussed alternative is the Bayesian statistical methodology, which, instead of generating a p value, generates a “posterior probability,” i.e., the final output of a Bayesian analysis is generated based, in part, on the consideration of a “prior probability” distribution of a random variable. This prior probability is subjectively assigned. Different posterior probabilities will be generated depending upon the choice of the prior probability used in the calculations; because the prior probabilities are subjective, this has contributed to the fact that these methods have not been widely adopted in the scientific community or in the regulatory environment. Another statistical methodology, a likelihood approach, has also been promoted as a methodology that its proponents assert answers the real question of interest in clinical trials: “When is a given set of observations evidence supporting one hypothesized probability distribution over another?”² This approach involves the explicit comparison of competing hypotheses on a given dataset, and its proponents argue that it lacks the theoretical disadvantages of either the frequentist or Bayesian approaches. It is fair to say that none of these methods is without flaw.

The approval of drugs to treat neurologic and psychiatric diseases incorporate all of the considerations dis-

cussed so far. Perhaps the most important consideration is the choice of the control group in these studies.

As noted above, the regulations describe the five types of controls that may serve to define a clinical study as adequate and well-controlled. In almost all cases of neurologic and psychiatric drug approval, a placebo concurrent control group has been used.

It is important to state at this point perhaps the most critical principle of clinical trial analysis routinely applied in the approval of neurologic or psychiatric treatments; namely, that the only trials that can be unambiguously interpreted to provide evidence of effectiveness of a new treatment are those in which a difference between active treatment and the control group is detected (operationally, again, this difference is defined as a statistically significant difference between the groups).^{3,4} An understanding of this principle is critical to an understanding of how the Agency interprets the results of clinical trials.

This principle is best explained in the context of an active controlled trial, i.e., a trial in which patients are randomized to receive either the new drug or an alternative, "active" treatment (usually a drug approved for the same indication). Typically, such trials are designed to demonstrate that the new and old treatments are "equivalent." Such a finding is taken to imply that the new treatment is effective, because it is "equal" to the old, effective treatment.

However, such a conclusion is based on an unstated assumption; namely, that in the specific trial discussed, the old treatment is effective. While this assumption appears perfectly reasonable on the surface, it may, in any instant case, be wrong. While the active comparator may be known to be effective (based on previously conducted controlled trials), this does not establish or imply that it was effective in the trial at hand. There are numerous examples of cases in which drugs known to be effective are not distinguished from placebo in particular trials. To interpret a trial that does not distinguish a new drug from an active comparator drug, one has to know that the patients assigned to the old drug would have been different (i.e., worse) had they not been treated with the old drug in this particular experiment. The only way this can be known with an acceptable degree of certainty is to have a robust body of clinical trial data in which the old drug essentially has always been found to be effective; that is, we would need to know that in all cases, the old drug is effective, so that we could be sure that, in the instant case, it was also effective. Unfortunately, for any specific drug to treat neurologic and psychiatric disease, this robust clinical trial data does not usually exist.

Active controlled trials that fail to demonstrate a difference between treatments are best viewed, therefore, as a type of historical controlled trial, because they rely on information external to the trial for their interpretation.

Although, as we have seen, the regulations consider a historical controlled trial as a potential adequate and well-controlled trial design, this design is only acceptable when the natural history of the untreated condition (which would serve as the historical control with which to compare the treated group) is known with great precision; unfortunately, for the conditions dealt with here, this information is usually not available.

Even if such a robust clinical trial database did exist, quantitative considerations make active controlled trials difficult, because a sufficient number of patients would need to be enrolled in the trial in order for it to be interpretable. That is, the null hypothesis in these trials (referred to as "non-inferiority" trials) is usually set up to be that the new treatment is no more than a certain amount worse than the active treatment (the so-called non-inferiority "margin"), and the study is designed to reject this new null hypothesis (the margin is usually chosen to represent the smallest difference seen between the comparator and control determined from the robust clinical trial experience with that treatment).⁵ Rejecting these sorts of null hypotheses ordinarily requires very large sample sizes.

Because interpreting active controlled trials that do not demonstrate a between-treatment difference is problematic for the reasons stated above, trials that do demonstrate a difference between treatments are preferred, and, indeed, required. For this reason, the placebo-controlled trial is the most efficient trial design, i.e., for an effective treatment, the likelihood that a difference will be seen between treatment and control is greatest with a placebo (inactive) treatment than with any other potentially active treatment (alternatively, an active controlled trial in which the new drug is statistically significantly superior to the control is acceptable as a positive study, assuming that the active control does not make the patients worse than they would have been with no treatment).

The use of placebo controls has come under fire on occasion, primarily because some consider its use unethical; this is particularly the case when other treatments are available for a given indication. However, the Agency's view has been that the use of placebo is appropriate and acceptable when withholding other available treatments does not result in permanent harm to the patient. That is, withholding treatments known to effect the underlying pathology of a disease are considered unethical (because withholding such treatment would result in a permanent deterioration), but, as a general matter, withholding symptomatic treatments (for an appropriate duration), is considered acceptable. This principle is endorsed by the Declaration of Helsinki⁶ and by the International Conference on Harmonization, which states:

"...when there is no serious harm, it is generally considered ethical to ask patients to participate in a placebo-

controlled trial, even if they may experience discomfort as a result, provided the setting is non-coercive and patients are fully informed about available therapies and the consequences of delaying treatment.”⁷

The placebo-controlled trials that the Agency relies on to support a finding of substantial evidence of effectiveness vary considerably in duration. Treatments designed to treat acute migraine headache typically study only a single headache (because subsequent headaches are intermittent and considered independent events that are expected to respond equally to treatment), while treatments for various forms of multiple sclerosis may have controlled portions of up to 2 years. As a general principle, studies of treatments for chronic conditions should (and do) assess drug effect for at least 3-6 months in a controlled setting, but the duration may need to be longer if the number of events necessary to be able to detect a drug-placebo difference are relatively rare (for example, studies of relapsing-remitting multiple sclerosis (MS) may need to be 1-2 years long in order for enough relapses to occur to be able to detect a between-treatment difference). Some studies of the acute phase of relatively chronic conditions [e.g., acute major depressive disorder (MDD), acute schizophrenia] are relatively brief (2-8 weeks), largely because patients cannot continue placebo treatment for very long (there are too many discontinuations after this duration to be able to adequately analyze the trial). In all instances in which prolonged treatment with placebo is untenable or clinically questionable, protocols contain contingencies to “rescue” patients with effective treatments.

In these latter conditions (chronic illnesses for which the controlled trials, especially in the acute phases, cannot be appropriately long), additional requirements are imposed on sponsors to document that the treatment is effective in the relatively long term. When placebo cannot be used over long periods (again, this usually applies in the setting of an acute phase of a more chronic disease), long-term efficacy is usually demonstrated in so-called randomized withdrawal designs.

In these studies, patients are treated with the drug in question in an open-label, uncontrolled setting. Those patients who are considered responders (by an agreed-upon definition), are continued on open-label drug for, ideally, at least 6 months, at which point they are randomized to continue on their treatment or placebo. The time to reaching “failure” criteria in this controlled portion of the trial is used to analyze the data, but the duration of drug effect (assuming that in the controlled portion the drug is shown to be superior to placebo) is derived from the open-label phase; this is why the goal is to have that phase be as long as practical. While in the past these longer duration studies were permitted to be completed after approval, the Agency has moved to re-

quiring these studies be done before marketing for those drugs in early development.

Another critical element of adequate clinical trials is a full exploration of the dose-response of the applied treatment. Specifically, a given (effective) dose will, on average, confer a particular “degree” of effectiveness, as well as adverse events. A full exploration of the range of effective doses, along with a characterization of the incidence and types of adverse events associated with these doses, permits the prescriber to make an informed decision about which dose (if any) may be appropriate for a particular patient. A full characterization of the dose-response should include the determination of the least effective dose as well as the maximum effective (and tolerated dose). An adequate dose-response can usually be characterized only in those studies in which patients are randomized to fixed doses (these final fixed doses may be achieved through titration). Many trials (particularly for psychiatric diseases) use so-called “flexible dose” designs, in which patients may be treated with doses within certain pre-specified ranges, based on the clinician’s judgment about their response. Because such trials are incapable of yielding any useful dose-response data (since patients are not randomized to particular doses their final dose is determined by many factors, including clinical response; in such cases, their response can be considered to “cause” the dose, rather than the reverse), they are discouraged. These designs make it difficult to offer useful prescribing information, because it is not clear which doses in the studied range are truly effective. Indeed, current guidelines suggest that the Agency has the authority to require adequate dose-response information before drug approval.⁸

An issue of great interest relates to the ability of clinical trials to yield useful comparative data. Specifically, many sponsors are interested in claiming that their product provides an advantage over other relevant treatments, either with regard to effectiveness and/or safety. Establishing either of these advantages is difficult.

The primary principle that needs to be considered in an assessment of trials designed to demonstrate an advantage of one treatment over another is that the comparison must be “fair.” That is, if a sponsor wishes to claim that their drug is more effective than another drug, the trials designed to demonstrate this superiority must compare relevant doses of the two drugs. Ideally, such a study would include the full dose range of both drugs, to determine the comparative efficacy across all appropriate doses. Similarly, a study designed to demonstrate the better tolerability of one drug compared to another not only would need to demonstrate a comparative benefit on the adverse event of interest, but might need to examine comparative effectiveness as well. Specifically, information that one drug is “safer” than another (with regard to the specific adverse event) is meaningful only if the

prescriber is aware of the relative efficacy at the doses of interest (i.e., to make an informed decision about which drug and dose to prescribe, the relative adverse event profile is meaningful only in the context of information about relative effectiveness). Further, such trials ordinarily focus on a single adverse event, but while a treatment may have a decreased risk for a particular adverse event compared to another drug, it may have a greater risk than that second drug for a different, or many different, adverse events. A comparative statement in labeling that refers only to the specific adverse event studied may be considered misleading if it does not describe the (perhaps increased) risk of other adverse events. Clearly, establishing comparative claims of this sort is difficult and, in general, these claims have not been granted for treatments for neurologic or psychiatric disease.

Considerable discussion has surrounded the question of the treatment effect size necessary to be detected in clinical trials to establish substantial evidence of effectiveness. For treatments of neurologic and psychiatric disease, there are no predetermined treatment effect sizes established. In the vast majority of cases, a statistically significant between-treatment difference on a face-valid measure of clinical signs and/or symptoms is sufficient to establish effectiveness. In most cases, these valid measures are rating scales that yield continuous values and are considered to reflect clinically meaningful functions. Examples include rating scales of symptoms and signs of Parkinson's disease, MDD, and schizophrenia. In other cases, counts of events are assessed. Examples include panic attacks and seizures. In all cases, the chosen outcome measure must assess at least some of what are considered the core symptoms/signs of a disease to support a labeling claim for that condition. For example, a global measure of how a patient "feels" would not typically be considered sufficient to establish effectiveness in MDD (or any other condition). For a specific claim to be granted to a treatment as an antidepressant, for example, a change on a measure that assesses symptoms and signs specific to depression must be obtained.

As noted earlier, in some settings (for example, treatments for Alzheimer's disease), two outcome measures are required to assess the clinical meaning of any changes seen on the measure of the core symptoms (in this case, a measure of cognitive functioning). This approach is being taken with more conditions recently but, as a general matter, statistically significant differences on a single appropriate clinical measure are considered evidence that the treatment is clinically useful.

An argument can be made that, instead of relying on mean changes on sensitive measures of signs and symptoms, response criteria should be established, and treatments should be compared on the basis of how many patients achieve responder status. While this approach is valid, the choice of the responder criteria is arbitrary and

always arguable. For example, in the past the Agency required that anti-seizure drugs be assessed by comparing the proportion of patients in both treatment groups who achieved a decrease in seizure frequency of at least 50% compared to their baseline. In these trials, patients who achieved a decrease in seizure frequency of 49% were not considered responders, although most would agree that, from a clinical point of view, there is no difference between this response and a 50% response. Because any choice of responder criteria will be arbitrary and personal, the Agency generally relies on mean changes on scales that measure signs and symptoms that are meaningful to the patient. Indeed, it is generally true that patients need not improve on treatment in order for a study to contribute to a finding of substantial evidence of effectiveness; in a given case, patients assigned to treatment may simply deteriorate less than patients assigned to placebo. What is critical is that a trial be able to detect a difference that is beneficial to patients (in this latter scenario, less worsening is considered beneficial) that is attributable to the treatment.

Of course, so many patients can be enrolled in a trial that miniscule differences between treatment and placebo can be detected to be statistically significant. In reality, while the definition of "miniscule" is open to interpretation, as a practical matter this does not occur. In any event, in most trials in which a mean difference is detected to be statistically significant, more patients assigned to drug achieve a given degree of improvement than placebo patients across the entire range of possible degrees of improvement (this is reflected in the construction of so-called cumulative distribution functions; see, for example, approved product labeling for any currently available treatment for Alzheimer's disease for an example of such a distribution).

An additional important element of trials submitted to support effectiveness is the geographical distribution of study sites. Multi-center studies are increasingly being conducted in both domestic and foreign institutions; many such trials include both foreign and domestic centers. There are examples, for a given drug, in which studies performed in the United States do not yield positive results when foreign studies do, as well as examples in which, in a given study, the results at domestic centers differ substantially from those seen in foreign centers. Whether the reasons for these discrepancies may be obvious (for example, the availability of different concomitant medications, different standards of care, different racial or genetic responses, etc.) or obscure, they may be critical to an adequate assessment of the effects of the drug in the US population. For this reason, although the regulations permit an application to be approved on the basis of entirely foreign data, other guidance gives the Agency the authority to require domestic studies if they are deemed necessary.⁹

Currently approved therapies to treat neurologic and psychiatric disease are considered to provide symptomatic benefit. Trials typically submitted to the Agency to support approval are not designed to demonstrate an effect on the underlying progression of any condition, and, if positive, therefore are interpreted as providing a symptomatic benefit. As noted earlier, the Act requires only that a drug be shown to have the effect claimed for it in labeling; as a result, it follows that the finding of a symptomatic benefit is quite sufficient to support drug approval (assuming the effect can be described adequately in product labeling). However, newer compounds are being developed that many consider to be potentially capable of inducing beneficial structural changes on the underlying pathology of many conditions. Two approaches have been proposed to address this question.

The first proposal consists of a clinical trial design that has been termed the randomized withdrawal design. In this trial, patients are randomized to treatment or control (again, invariably placebo) as in a typical symptomatic study (Part 1). After the appropriate duration and when a difference has been shown between the treatments, patients originally randomized to treatment are switched to treatment with placebo, while the patients originally treated with placebo are continued on placebo (part 2). After an appropriate duration, the original drug-treated patients (now on placebo) are compared to the original placebo patients (continued on placebo). If the response in the former group has approached that of the latter group, a symptomatic effect is implied (that is, this result demonstrates that having been treated in the initial period did not prevent these patients from ending up where they would have been had they not been treated). However, if the original drug-treated group maintains the same difference from the placebo patients at the end of part 2 as existed at the end of part 1, this implies a structural effect (i.e., having been treated early implies that these patients are different than they would have been had they not been treated). Such trials pose practical problems (for example, the duration of the withdrawal phase may need to be quite long and, for statistical reasons, the trial may need to enroll large numbers of patients) but if well designed and conducted, they can be interpreted unambiguously as demonstrating a structural effect.¹⁰ Unfortunately, these trials are rarely performed.

More recently, there has been considerable interest in relying on the effect of a drug on various surrogate markers to establish its effect on the underlying structural elements giving rise to disease.

It is important to note at this point that while there have been many candidate surrogate markers proposed to support this use (e.g., various imaging and biochemical measures in various conditions), there is general agreement that none of these measures constitute a valid sur-

rogate marker, as defined earlier, for any neurologic or psychiatric disease. However, as noted earlier, the Act and regulations permit the Agency to approve a drug on the basis of its effect on an unvalidated surrogate marker. To date, however, the Agency has not done so for neurologic or psychiatric treatments, for several reasons. First, as discussed above, it is not necessary for drug approval. While it is possible to imagine that a drug that induces a beneficial structural effect may not induce an otherwise clinically detectable effect in a reasonable period of time (one proposed advantage of surrogate markers as primary outcomes is that they may show changes fairly quickly), for the treatments so far approved, clinically detectable effects can be demonstrated with reasonable sample sizes in reasonable periods of time (ranging from a single dose for the acute treatment of migraine, to 3-6 months for treatments for Alzheimer's disease, to up to 2 years for treatments for MS). While some of the currently approved treatments might, in fact, induce beneficial structural changes, such a demonstration was, and is, clearly unnecessary to obtain approval.

The primary reason, however, that the Agency has not relied on a drug's effect on an unvalidated surrogate marker to support approval is that relying on such an effect may be misleading.

Specifically, for an unvalidated surrogate, we cannot be certain that a beneficial effect seen on the surrogate will translate into the desired clinical benefit (by definition). While most candidate surrogates correlate very well with disease progression (e.g., hippocampal atrophy as seen with magnetic resonance imaging and progression of Alzheimer's disease) in the untreated state, there is no guarantee that a beneficial change on the surrogate with treatment will definitively result in the desired clinical change. There are numerous reasons why this may be so, perhaps most importantly because drugs have many effects that are unpredictable, and they may induce deleterious effects on the desired clinical outcome in addition to a beneficial effect on the surrogate. Numerous examples exist in which the applied treatment had the desired effect on the surrogate, but a negative (or no) effect on the desired clinical outcome. These issues, including examples of "failed" surrogates and, more importantly, the reasons why surrogates may fail, were discussed in detail previously.¹¹

What is important to note here, though, is that we cannot have great confidence in relying on a drug's effect on an unvalidated surrogate unless we have almost complete knowledge of the pathophysiology of the disease being treated, and both the positive and negative mechanisms of action of the applied treatment.¹¹ This knowledge is not available for any neurologic or psychiatric condition with which the Agency deals, nor is it available for any of the treatments approved (or regulated) by the Agency. For this reason, approving a drug on the basis of

its effect on an unvalidated surrogate marker, although possible and appropriate at times, requires making numerous assumptions not ordinarily necessary in a typical case of drug approval.

Although there is considerable information available both on the (presumed) mechanism of action of the drug under review, and the pathophysiology of disease under study, this information is typically secondary in drug approval. The Agency's primary regulatory philosophy can best be characterized as empiricist, one formal definition of which is the "...thesis that all knowledge of non-analytic [non-definitional] truths...is justified by experience."¹² Simply put, the Agency adopts an empirical approach to the fundamental regulatory questions of safety and effectiveness.

Theories about mechanism of action of a drug or disease mechanisms play important parts in drug development and approval, but they are entirely subsidiary to the fundamental questions that must be answered in the course of drug approval; namely, is a drug effective, and is it safe in use. These questions can only be answered (within the limits of experimental error and statistical uncertainty) by direct examination of the question in a well designed and conducted clinical experiment. These conclusions cannot, in the typical case, be predicted, nor can they be arrived at by an "understanding" of the underlying events, an understanding that must always remain incomplete (and, importantly, incomplete in ways that are unknown to us). For this reason, the Agency endeavors to conclude that a drug is effective by relying on as few assumptions as possible. This is best accomplished by relying on the results of adequately designed

and conducted clinical experiments that permit the direct determination that a treatment is effective on signs and/or symptoms of concern to patients.

REFERENCES

1. Guidance for industry—Providing evidence of effectiveness for human drug and biological products. U.S. Department of Health and Human Services, May 1998.
2. Royall R. Statistical evidence: a likelihood paradigm. In: Monographs on statistics and applied probability (Cox DR, Isham V, Keiding N, Reid N, Tong H, eds), p 31. London: Chapman & Hall, 1997.
3. Leber P. Hazards of inference: the active control investigation. *Epilepsia* 30 [Suppl 1]:S57–S63, 1989.
4. Temple R, Ellenberg S. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Ann Intern Med* 133:455–463, 2000.
5. Temple R, Ellenberg S. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Ann Intern Med* 133:464–470, 2000.
6. World Medical Association. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull World Health Organ* 79:373–374, 2001.
7. Guidance for industry: E10 choice of control group and related issues in clinical trials. US Department of Health and Human Services, May 2001.
8. Guideline for industry: E4 dose-response information to support drug registration. U.S. Department of Health and Human Services, November 1994.
9. Guideline for industry: E5 ethnic factors in the acceptability of foreign clinical data. U.S. Department of Health and Human Services, 1998.
10. Leber P. Slowing the progression of Alzheimer disease: methodologic issues. *Alzheimer Dis Assoc Disord* 11 [Suppl 5]:S10–S21, 1997.
11. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* 125:605–613, 1996.
12. Rosenberg A. Philosophy of science: a contemporary introduction (Moser PK, ed). Routledge contemporary introductions to philosophy. London: Routledge, 2000.