CONTINUING MEDICAL EDUCATION

Understanding and evaluating clinical trials

Michael Bigby, MD, a and Anne-Sophie Gadenne, MD b Boston, Massachusetts, and Cincinnati, Ohio

In this review, we attempt to provide the basic knowledge necessary to understand and evaluate clinical trials because properly conducted, randomized clinical trials are the best sources for determining the best available treatment. Other commonly used sources rarely provide sufficient detail necessary to determine the efficacy and safety of any treatment, and they often contain biases or pitfalls that make them unacceptable or unreliable. Our method for reviewing clinical trials allows a busy clinician to use his or her time most efficiently by deciding not to read the majority of poorly conceived, designed, executed, or reported trials and those trials with insignificant results. It provides a means to determine the quality of the trials that one does decide to read and to retain and retrieve the information when it is needed. The method involves recognizing and evaluating the features that strengthen clinical trials and help validate their conclusions. These features include proper selection and allocation of patients, inclusion of an appropriate control group, randomization, prior selection of clinically and biologically important outcome variables, blinding of assessment, consideration of patient compliance and drop out, and proper presentation of statistical analysis of results. (J AM ACAD DERMATOL 1996;34:555-90.)

Learning objective: At the conclusion of this learning activity, participants should be able to understand and critically evaluate clinical trials.

This review attempts to provide the knowledge necessary to understand and evaluate clinical trials because they are the best sources for determining the best available treatment. The review is based on several premises. The first is that clinical trials published in peer-reviewed journals are the best sources and are readily available. They can be found by using one of several software interfaces for searching the MEDLINE database.

However, published clinical trials should also be approached with a sense of skepticism. In a review of more than 30 years of clinical trials, Hemminki found that the majority were uncontrolled, poorly controlled, or otherwise poor. Only 18% to 27% were judged to be "good" in the years 1943 to 1974. Reviews of more recent clinical trials have yielded similar results. There are many practical and logistical difficulties inherent in conducting therapeutic trials in human subjects. Adherence to all recommended methods is not always possible; therefore it is impossible to conduct the "perfect" clinical trial. However, the third premise is that investigators must provide readers with adequate information about the methods employed in published clinical trials. The fourth premise is that physicians currently try therapies and continue to use therapies whose use is based on poorly conducted clinical trials or no trials at all.
METHODS FOR ASSESSING CLINICAL TRIALS

Many strategies for critically reviewing therapeutic trials have been developed and have been well reviewed.27 The method that we discuss is a modification of the method of DerSimonian that includes an assessment of the trials' results and the use of 15 methodologic items.19,20 The 15 methodologic criteria are shown in Table I. Each is described and discussed in turn. The first five deal with selection and allocation of patients.

CRITERIA FOR ASSESSING CLINICAL TRIALS

Eligibility criteria

The first step in specifying eligibility criteria is to state explicitly the diagnostic criteria for the disease being studied. For example, the diagnostic criteria for acne or psoriasis can be simply stated. However, in some instances, defining diagnostic criteria is more difficult, requires more precision, and is of critical importance. For example, in a study of cutaneous T-cell lymphoma, the diagnostic criteria need to be defined precisely because both the pathologic and clinical definitions are controversial.28-32

The second step is to define the specific population of patients being studied. Were all patients who met the diagnostic criteria eligible for the study? If not, which subsets were included and which subsets excluded? The demographics of the patient population eligible for inclusion should be described. It is fundamental to recognize that results obtained in studies of skewed populations (e.g., patients unresponsive to treatment or patients who have been referred to tertiary care centers) may not be directly applicable to general practice. Conversely, results obtained from studies of patients who have not had prior treatment may not be directly applicable to patients who are referred to tertiary care centers because they have failed to respond to standard forms of therapy.

When faced with the task of determining whether the results of a particular study are applicable to a specific patient, physicians should ask themselves whether there are any compelling reasons that the results should not be applied to the patient. Only if the answer is yes are the study's results not applicable to the patient.21,33

Admission before allocation

To eliminate either conscious or unconscious biases, admission or exclusion should occur before patients are allocated to treatment groups. Admission of patients to studies after they have been allocated to a treatment group is a source of unrecognized and underreported biases in clinical trials.19,20 It allows the selection of known or anticipated responders (or nonresponders) to treatment groups in a biased fashion. Consider the following example:

A physician becomes interested in a new topical therapy for tinea pedis after observing that in the first four patients he treats with it the skin clears within 2 weeks. He then treats an additional six patients with the new cream and finds that three have a similar response. As a control group, he treats the next 10 consecutive patients with the cream he formerly used most commonly and finds that in only two of them does the skin clear within 2 weeks. The cure rate of the new cream (70%) seems clearly superior to the cure rate of the old cream (20%). However, by admitting patients to his study after allocating them to a treatment group (and observing their positive response to treatment), the physician has introduced a serious bias in favor of the new cream. The new cream may in fact be superior, but that conclusion cannot be drawn justifiably from the data.

Control groups

A control group that is studied concurrently is an essential component of most adequate clinical trials.13,34 The utility of a control group was well illustrated by a survey of psychotropic drug studies in which 83% of 52 investigations without controls 9.14
indicated that a drug was effective whereas only 25% of 20 studies using controls indicated effectiveness.\(^{13,35}\)

The benefits or side effects that occur during the course of a clinical trial may result from many factors, only one of which is the effect of treatment.\(^{33,36}\) Other factors that may influence outcome include the natural history of the disease, the nonspecific effects of treatment, chance, and bias.\(^{33}\) Many conditions are characterized by waxing and waning. Because patients are most likely to seek treatment or enter a study when they are doing poorly, the natural history of many diseases makes it likely that they would often improve after entry whether therapy is effective or not.\(^{36}\) Outcome is also influenced by the attention received during a trial and by patients’ and physicians’ expectations of treatment.\(^{33}\) These nonspecific factors are referred to as placebo effects.\(^{13,36}\) Placebo effects can range from 15% to 58%.\(^{36,37}\) Placebos are also associated with side effects in approximately 19% of patients, as documented in a review of 109 controlled trials.\(^{36,38}\) A concurrently studied control group is the best way to ensure that the results obtained are from the treatment and not from the natural history of the disease or placebo effects.\(^{36}\)

The only clinical trials in which no controls are necessary are those studying the treatment of diseases that are universally and rapidly fatal in which any survival will be the result of the treatment or conditions in which the outcome is uniform and well documented.\(^{33,39}\) For example, the efficacy of ivermectin in the treatment of scabies in healthy and HIV-infected patients was strongly suggested by an open-labeled, uncontrolled clinical trial.\(^{40}\)

The choice of an appropriate control group is a critical issue in the design of clinical trials.\(^{41}\) Placebo controls are appropriate for diseases for which there is no proven effective treatment. A placebo or untreated control group may not be appropriate for diseases for which there is proven effective treatment or a well-established standard treatment.\(^{41,42}\) In these situations, groups treated with the proven or standard therapy, at appropriate doses, would be the most appropriate controls.\(^{1,43}\) Placebo controls also may be inappropriate for fatal diseases on both ethical and logical grounds.\(^{41}\)

Whereas it is tempting to use “historic controls,” especially for uncommon conditions, making unbiased comparisons between current and previous experiences is extremely difficult but not impossi-

ble.\(^{13,33,39,44,45}\) Among the problems of historical controls are their insensitivity to differences in selection criteria and changes in ancillary care.\(^{44}\) These problems have been well documented and substantiated even in the treatment of cancer.\(^{33,34,44,46}\) For example, in a review of 20 studies of fluorouracil treatment of advanced cancer of the colon, Moertel and Reitermeier\(^ {47}\) found that response rates ranged from 8% to 85%. Thus the same treatment of the same disease produced widely different response rates in different studies at different times.\(^{33,47}\) However, studies utilizing historic controls have played important roles in clinical investigation.\(^{32}\) For example, interferon alfa was introduced and won approval by the Food and Drug Administration (FDA) for the treatment of hairy cell leukemia on the basis of small, historically controlled trials.\(^{48-54}\)

Obtaining adequate controls is a common problem. Many conditions are uncommon and recruiting adequate numbers of patients is difficult. One approach is to utilize multiple large centers to recruit an adequate number of patients. This approach may lead to important differences in protocol or patient selection that may result in significantly different responses in different centers. For example, in a multicenter trial of photopheresis for the treatment of cutaneous T-cell lymphoma, one center had an inexplicably low favorable response rate of 29% (2 of 7) in contrast to the remaining four centers that had a combined favorable response rate of 83% (25 of 30).\(^ {55}\)

Another approach is to use patients as their own controls either in crossover studies or in split studies in which half the body is treated one way and the other half is treated another. Patients may serve as their own control in studies designed to compare untreated periods to periods of treatment. These approaches have important limitations.\(^ {13,45,56}\)

In crossover studies, it is particularly important to pay attention to treatment order and rules for the timing of crossover.\(^ {13,56}\) Ideally, patients in crossover studies should be randomly assigned to receive any of the available treatments as the first treatment. In other words, nearly equal numbers of patients should receive the new treatment as the first treatment or the comparison treatment (or placebo) as the first treatment. The point at which patients are “crossed over” from one treatment to another may be a source for error or the introduction of bias. Ideally, the crossover point should be at a specific pre-
indicated that a drug was effective whereas only 25% of 20 studies using controls indicated effectiveness.\textsuperscript{13, 35}

The benefits or side effects that occur during the course of a clinical trial may result from many factors, only one of which is the effect of treatment.\textsuperscript{33, 36} Other factors that may influence outcome include the natural history of the disease, the nonspecific effects of treatment, chance, and bias.\textsuperscript{33} Many conditions are characterized by waxing and waning. Because patients are most likely to seek treatment or enter a study when they are doing poorly, the natural history of many diseases makes it likely that they would often improve after entry whether therapy is effective or not.\textsuperscript{36} Outcome is also influenced by the attention received during a trial and by patients’ and physicians’ expectations of treatment.\textsuperscript{33} These nonspecific factors are referred to as placebo effects.\textsuperscript{13, 36} Placebo effects can range from 15\% to 58\%.\textsuperscript{36, 37} Placebos are also associated with side effects in approximately 19\% of patients, as documented in a review of 109 controlled trials.\textsuperscript{36, 38} A concurrently studied control group is the best way to ensure that the results obtained are from the treatment and not from the natural history of the disease or placebo effects.\textsuperscript{36}

The only clinical trials in which no controls are necessary are those studying the treatment of diseases that are universally and rapidly fatal in which any survival will be the result of the treatment or conditions in which the outcome is uniform and well documented.\textsuperscript{33, 39} For example, the efficacy of ivermectin in the treatment of scabies in healthy and HIV-infected patients was strongly suggested by an open-labeled, uncontrolled clinical trial.\textsuperscript{40}

The choice of an appropriate control group is a critical issue in the design of clinical trials.\textsuperscript{41} Placebo controls are appropriate for diseases for which there is no proven effective treatment. A placebo or untreated control group may not be appropriate for diseases for which there is proven effective treatment or a well-established standard treatment.\textsuperscript{41, 42} In these situations, groups treated with the proven or standard therapy, at appropriate doses, would be the most appropriate controls.\textsuperscript{1, 45} Placebo controls also may be inappropriate for fatal diseases on both ethical and logical grounds.\textsuperscript{41}

Whereas it is tempting to use "historic controls," especially for uncommon conditions, making unbiased comparisons between current and previous experiences is extremely difficult but not impossi-

ble.\textsuperscript{13, 33, 39, 44, 45} Among the problems of historical controls are their insensitivity to differences in selection criteria and changes in ancillary care.\textsuperscript{44} These problems have been well documented and substantiated even in the treatment of cancer.\textsuperscript{3, 53, 44, 46} For example, in a review of 20 studies of fluorouracil treatment of advanced cancer of the colon, Moertel and Reitemeier\textsuperscript{47} found that response rates ranged from 8\% to 85\%. Thus the same treatment of the same disease produced widely different response rates in different studies at different times.\textsuperscript{33, 47} However, studies utilizing historic controls have played important roles in clinical investigation.\textsuperscript{34} For example, interferon alfa was introduced and won approval by the Food and Drug Administration (FDA) for the treatment of hairy cell leukemia on the basis of small, historically controlled trials.\textsuperscript{48-54}

Obtaining adequate controls is a common problem. Many conditions are uncommon and recruiting adequate numbers of patients is difficult. One approach is to utilize multiple large centers to recruit an adequate number of patients. This approach may lead to important differences in protocol or patient selection that may result in significantly different responses in different centers. For example, in a multicenter trial of photopheresis for the treatment of cutaneous T-cell lymphoma, one center had an inexplicably low favorable response rate of 29\% (2 of 7) in contrast to the remaining four centers that had a combined favorable response rate of 83\% (25 of 30).\textsuperscript{55}

Another approach is to use patients as their own controls either in crossover studies or in split studies in which half the body is treated one way and the other half is treated another. Patients may serve as their own control in studies designed to compare untreated periods to periods of treatment. These approaches have important limitations.\textsuperscript{13, 45, 56}

In crossover studies, it is particularly important to pay attention to treatment order and rules for the timing of crossover.\textsuperscript{13, 56} Ideally, patients in crossover studies should be randomly assigned to receive any of the available treatments as the first treatment. In other words, nearly equal numbers of patients should receive the new treatment as the first treatment or the comparison treatment (or placebo) as the first treatment. The point at which patients are "crossed over" from one treatment to another may be a source for error or the introduction of bias. Ideally, the crossover point should be at a specific pre-
determined time (e.g., after 10 weeks of the first treatment period) and both the patient and the assessor of outcome should be blinded as to when crossover occurs. If the crossover point is determined by the patient’s response to treatment, significant biases may be introduced that favor one treatment over another.\(^{13,56}\)

For example, consider the results that would be obtained if a disease is known to wax and wane, and a crossover study is designed to shift patients who are doing poorly from placebo to experimental drug but is not designed to shift patients who are doing poorly with an active drug to placebo. Often such studies are designed to adjust the dose of patients taking active drug who are doing poorly. Some patients receiving placebo who are doing poorly would naturally be shifting from doing poorly to doing better. They will appear to improve when the switch to active drug is made whether the drug is effective or not (Fig. 1). Some patients who are doing poorly with active drug would naturally be shifting from doing poorly to doing well. They too will appear to improve after their dose is adjusted. The result of this study design is to produce the appearance that a treatment is effective even if it is not and to exaggerate the magnitude of the effect of the drug if it is effective.\(^{13}\)

There are other important caveats to consider. First, crossover studies should be limited to diseases in which the treatment is short-acting and the patient’s condition is expected to revert to its original state once treatment effects are gone.\(^{13,56,57}\) If a treatment has prolonged effects, a sufficient wash-out period, when the patient receives no treatment, is essential to allow the effects of the first treatment to dissipate.\(^{13,56,57}\) Finally, loss of patients in small crossover studies may lead to large changes in the validity and strength of the statistical results.\(^{56}\)

Special care must be taken in designing bilateral comparison trials of topical therapies. Topically applied preparations are easily translocated from one site to another, as was demonstrated with fluorescence used to identify the unintentional translocation of topical tetracycline.\(^{58}\) Bilateral designs also can be difficult to interpret because of systemic absorption and systemic effects. These effects are sometimes totally unexpected. For example, in a bilateral comparison study of lichen planus, in some patients treated on one side with PUVA the skin cleared on both sides.\(^{59}\)

**Random allocation**

Ideally, patients should be randomly assigned to treatment groups to avoid the introduction of biases and to distribute intrinsic prognostic factors equally among treatment groups.\(^{33,39,40,41,45,60}\) Randomization controls patient selection and evaluation bias by physicians and study personnel.\(^{33,61,62}\)
obtained if a disease is known to wax and wane, and a crossover study is designed to shift patients who are doing poorly from placebo to experimental drug but is not designed to shift patients who are doing poorly with an active drug to placebo. Often such studies are designed to adjust the dose of patients taking active drug who are doing poorly. Some patients receiving placebo who are doing poorly would naturally be shifting from doing poorly to doing better. They will appear to improve when the switch to active drug is made whether the drug is effective or not (Fig. 1). Some patients who are doing poorly with active drug would naturally be shifting from doing poorly to doing well. They too will appear to improve after their dose is adjusted. The result of this study design is to produce the appearance that a treatment is effective even if it is not and to exaggerate the magnitude of the effect of the drug if it is effective.13

There are other important caveats to consider. First, crossover studies should be limited to diseases in which the treatment is short-acting and the patient’s condition is expected to revert to its original state once treatment effects are gone.13, 36, 57 If a treatment has prolonged effects, a sufficient wash-out period, when the patient receives no treatment, is essential to allow the effects of the first treatment to dissipate.13, 36, 57 Finally, loss of patients in small crossover studies may lead to large changes in the validity and strength of the statistical results.56

Special care must be taken in designing bilateral comparison trials of topical therapies. Topically applied preparations are easily translocated from one site to others, as was demonstrated with florescence used to identify the unintentional translocation of topical tetracycline.58 Bilateral designs also can be difficult to interpret because of systemic absorption and systemic effects. These effects are sometimes totally unexpected. For example, in a bilateral comparison study of lichen planus, in some patients treated on one side with PUVA the skin cleared on both sides.59

Random allocation

Ideally, patients should be randomly assigned to treatment groups to avoid the introduction of biases and to distribute intrinsic prognostic factors equally among treatment groups.33, 39, 44, 45, 60 Randomization controls patient selection and evaluation bias by physicians and study personnel.33, 61, 62
Nonrandomized studies have limited value in distinguishing useful from useless or even harmful therapy.\textsuperscript{14, 22, 33, 45} Studies in which patients are not allocated randomly more commonly show larger effects, many of which are false positive, than randomized trials.\textsuperscript{22, 45} For example, the belief that azathioprine reduces the dose of corticosteroids necessary to treat bullous pemphigoid was suggested by several uncontrolled or nonrandomized studies.\textsuperscript{63, 64} However, a randomized trial of the effects of azathioprine on corticosteroid dose in patients with bullous pemphigoid indicated that azathioprine had no or negligible effects on remission or corticosteroid doses.\textsuperscript{65}

Randomization has the disadvantage of depending on luck for an equal statistical distribution of prognostic factors among groups.\textsuperscript{1, 66} Therefore it is always important to identify known prognostic factors and compare their distribution among treatment groups at baseline.\textsuperscript{22, 23, 33, 67, 68} The common practice of assigning \textit{p} values to differences in characteristics at baseline is inappropriate.\textsuperscript{23, 44, 45, 67-69} Demographic and known prognostic factors may not be equally distributed among treatment groups when patients are randomly assigned to treatment groups. However, the impact of the unequal distribution of prognostic factors on treatment results should be assessed.\textsuperscript{70-72} Inequality of groups in randomized trials is a more significant problem in studies with few patients.

Several allocation strategies have been employed to ensure that demographic and important known prognostic factors are equally distributed among treatment groups.\textsuperscript{57, 73} One is to pair or match patients in terms of demographic characteristics or disease features (e.g., type and severity of disease) and then randomly assign one member of each pair to one treatment and the partner to the other.\textsuperscript{57, 73} An extension of this strategy is the randomized block design. In this system, groups or blocks of patients with similar demographic or clinical features are entered into a study and then patients within the group are randomly assigned to alternative treatments in equal numbers. These techniques often strengthen the design of trials because they ensure that important, known prognostic factors are equally distributed among treatment groups, that nearly equal numbers of patients will receive each treatment, and that the advantages of randomization are maintained. The randomized block design is especially useful in multicenter trials.\textsuperscript{57, 73}

**Method of randomization**

When patients are assigned randomly, the method used to randomize should be reported. The most widely accepted and unbiased method is using sealed random numbers obtained from tables or pseudorandom numbers generated by computers.\textsuperscript{20, 45, 67} In this method, patients are assigned a random number after they are admitted to a trial. That number has been previously assigned to a treatment group and the assignment was unknown by the patient or the investigator. Correct randomization or concealment of assignment are often not performed in published studies, and the methods of randomization are rarely reported.\textsuperscript{19, 20, 67}

Many commonly encountered methods of randomization (e.g., on the basis of admitting team, alternative assignment, odd or even days of entry, birth dates, social security numbers, or day of the week) can lead to significant introduction of bias.\textsuperscript{20, 33, 39, 44, 45, 60, 67, 72} For example, a physician who would like a new treatment to be superior in efficacy to an older treatment may be able to have patients most likely to respond to treatment entered into the study on even days and patients least likely to respond on odd days if he or she knows that the new treatment is assigned on even days and the older treatment is assigned on odd days.

The next six criteria deal with the mechanisms involved in measuring the response to treatment.

**Blindness to treatment and blind assessment of outcome**

It is important to make the evaluation of the outcome of a trial as blinded as possible to avoid the introduction of bias.\textsuperscript{33, 39} "Blinding" of physicians and patients is most important when subjective outcomes are being measured.\textsuperscript{39} "Blinding" of patients and evaluators ensures that ancillary therapies and outcome evaluations are applied equally.\textsuperscript{33} A patient who is known to be receiving a new treatment may be observed more closely, may receive better ancillary care, and may have different expectations than a control patient. Changes in disease status may receive greater (or lesser) emphasis, and adverse events may similarly receive more (or less) attention.\textsuperscript{13, 33}

However, blinding is often not possible.\textsuperscript{13, 33} Many treatments produce skin changes that make it apparent what treatment is being used. For example, the drying effects of topical and systemic retinoids and the tanning effects of UVB therapy make the
recipients of those therapies readily apparent. An unblinded study may still be valid but the designers of an unblinded study must make allowances for the possibility that bias may have been introduced and should increase the statistical rigor for demonstrating differences.

Clearly defined outcome variables

Two principal methods are used to determine patient outcomes in dermatologic clinical trials. The first involves examining patients before, during, and at the conclusion of treatment, and reporting the appearance of the patients at the various time points. The second involves determining the degree of improvement during treatment. A third method, determining the impact of therapy on the quality of the patient’s life, is not commonly used in dermatologic trials.

An example of the first method is commonly encountered in therapeutic trials of psoriasis. A common practice is to assign numeric values to the amount of erythema, scaling, and degree of infiltration, to determine the area of the body surface involved, and to formulate an “index” by calculating a derivative of some product of these four numbers. The overall condition of the patient can then be represented by this index. A common example is the Psoriasis Area and Severity Index (PASI), which ranges from 0 to 72. The major problem with indices is that they confound area of involvement with severity of disease. For example, a patient with thick plaque-type psoriasis of the knees, elbows, and scalp may have the same index as a patient with diffuse but minimal psoriasis of the trunk and arms. Whereas the former patient is notoriously difficult to treat, the latter will generally respond rapidly and easily to many forms of therapy. The second problem with indices is that they lend an air of precision to the analysis and presentation of data that is not deserved. For example, Tilling-Grosse and Rees demonstrated that physicians and medical students were poor at estimating the area of skin disease and therefore some of the components that make up indices may be inaccurate. Finally, calculating the means, differences in means, and percentages of change in indices in response to treatment often do not convey an accurate clinical picture of the changes that have occurred.

The second method of assessment groups patients according to their degree of improvement. Treatments are then compared by their ability to move patients to higher degrees of improvement. There are two major problems with this form of assessment. The first is that the categories of improvement are often not well defined. The second problem is that the categories are not additive; that is, 60% to 80% improvement is often assumed to be twice as good as 20% to 40%, although no such numeric relation exists between these subjectively defined categories.

To be most useful, the outcome variables to measure must be clearly defined, as objective as possible, and have clinical and biologic significance. The best indices and scales are the ones that accurately reflect the state of the disease and the ones whose validity has been verified by previous work. The development of scales and indices for cutaneous diseases and testing their validity, reproducibility, and responsiveness have been inadequate. Therefore a lack of clearly defined and useful outcome variables remains a major problem in interpreting dermatologic clinical trials.

Until better scales are developed, trials with the simplest and most objective outcome variables are the best. They lead to the least amount of confusion and have the strongest conclusions. Thus trials in which a comparison is made between death and survival, patients with recurrence of disease and those without recurrence, or patients who are cured and those who are not cured are studies whose outcome variables are easily understood and verified. For trials in which the outcomes are less clear cut and more subjective in nature, a simple ordinal scale is probably the best choice. The best ordinal scales involve a minimum of human judgment, have a precision that is much smaller than the differences being sought, and are sufficiently standardized that they can be used by others and produce similar results.

In developing scales, authors should clearly define in words what the response categories are and what they mean. It is helpful to have precise descriptions accompanied by illustrative photographs. Stating that in response to treatment, a rash was “unchanged, improved, or cleared” is not clearly defining the response. Nor does assigning numeric values 0, 1, and 2 to unchanged, improved, and cleared, respectively, define the response.

In addition to being clearly defined, outcome variables should have clinical and biologic significance. For example, in a therapeutic trial of patients with severe acne, treatment was associated
with a decrease in lesion count from a mean of 400 to a mean of 350. This numeric difference may be of statistical significance, but it does not convey what the biologic significance of the change in lesion number represents. This result may mean that in some patients with severe acne the skin cleared completely, whereas in other patients the skin remained the same or got worse. It could also mean that most patients got slightly better. Furthermore, does an individual patient look better when their lesion number has been reduced from 400 to 350? Are scarring and complications less?

To strengthen clinical trials and help validate their conclusions, the outcome variables should be few and should be chosen before initiation of the study. Many outcome variables increase the likelihood that spurious, chance differences will be detected. An ineffective treatment may be found efficacious when tested with poorly designed outcome assessment tools. Conversely, an effective therapy may be found ineffective by an insensitive scale.

Special precautions are recommended in recognizing and being skeptical of "substitute or surrogate end points" especially when no differences are detected in clinically important outcomes. Examples of such end points include CD4/CD8 ratios instead of survival in studies of treatments of AIDS, antinuclear antibody levels or sedimentation rates instead of clinical measures of disease activity in lupus erythematosus, and volume of warts instead of proportion of patients cleared of warts. Carefully chosen and validated surrogate end points often allow smaller studies to provide answers to questions that would typically require much larger trials if the targeted clinical end point were utilized. For example, a well-designed, short clinical trial may be sufficient to demonstrate that a new drug effectively reduces the serum cholesterol or that another drug is effective in controlling hypertension. In both cases much longer and larger studies would be required to demonstrate that the cholesterol-reducing drug and the antihypertensive drug reduced morbidity and mortality from atherosclerotic and hypertensive cardiovascular diseases, respectively. Surrogate end points must, however, correlate with clinical outcomes, and their validity must be demonstrable in previous studies. A dramatic example of the dangers of substitute end points was provided by studies addressing the usefulness of antiarrhythmic drugs after myocardial infarction. The drugs were shown to reduce the occurrence of abnormal ventricular depolarization (a substitute end point) after myocardial infarction, but in randomized controlled trials, their use was associated with excess mortality when compared with placebo.

Compliance

Differences in outcome in clinical trials may be caused by differences in compliance and not by differences between the treatments if they were both used appropriately. Compliance will be lower for treatments that are difficult, messy, or have significant side effects. Methods for assessing compliance include questioning the patients, counting or weighing returned medication, or supervised treatment. Compliance is almost never addressed or reported in dermatologic therapeutic trials.

It is important not to exclude noncompliant patients from the analysis in the group to which they were randomly assigned. Analyzing even noncompliant patients with the group to which they were randomized is the basis of an "intention-to-treat analysis." It ensures that the prognostic factors that were evenly distributed by randomization are preserved.

Equally important is not to try to establish the efficacy of treatment by comparing compliant patient with noncompliant patients. The fallacy of this approach is illustrated by results from a study of men who survived a myocardial infarction who were observed prospectively for the development of repeat infarction or death. Twenty-six percent of 882 men who took less than 80% of their medication died compared with only 16% of the 1813 men who took more than 80% of their medication. This highly biologically and statistically significant difference seems to provide strong evidence that the medication was efficacious until we learn that the medication was placebo. The fact that compliant patients who follow instructions may have better outcomes whether their treatment works or not has been verified in several studies. This principle was evidently overlooked by the authors, reviewers, and editors of an article suggesting that dinitrochlorobenzene application improves survival in HIV-1-infected patients based on the observation that compliant and noncompliant patients appeared to differ in survival.

Complications

The nature and frequency of adverse events that occur during a clinical trial should be recorded and
reported, as should the responses made to these events (e.g., none, increased surveillance, change in treatment dose, counteractive treatment, or drug discontinuation).\textsuperscript{73} Adverse experiences (occurrences in a study that may or may not be related to a drug or treatment) and adverse effects of treatment (side effects that are attributable to the treatment) must be distinguished. The latter can be anticipated from known properties of the drug.\textsuperscript{73} They may also be distinguished by examining statistically significant differences in adverse experiences between placebo and active groups. Simple and sophisticated methods and algorithms for determining the strength of the causal relation between treatment exposure and the development of an adverse experience are available.\textsuperscript{73, 95-98} The previous experience with the drug, presence or absence of alternative explanations of the adverse experience, timing, dose-response relations, and consequences of withdrawal and rechallenge are taken into consideration in these assessments.\textsuperscript{73, 95-98}

Controlled clinical trials that typically enroll small numbers of patients are often not adequate for determining the long-term toxicity of newly introduced treatments.\textsuperscript{19, 69} For example, in premarketing controlled trials of benoxaprofen and zomepirac sodium, both drugs were shown to be efficacious and relatively safe; however, both drugs were withdrawn from the market shortly after they were marketed. Wide use revealed that benoxaprofen was associated with hepatic failure and death in elderly patients and zomepirac sodium was associated with anaphylaxis and death.\textsuperscript{100} Conversely, if a well-designed randomized controlled trial does indicate a clinically and statistically significant relation between a treatment and an adverse effect, the association is likely to be valid.\textsuperscript{99} Other study designs (e.g., cohort studies, case-control studies, case series, and case reports) may be necessary to assess the potential harmful effects of treatments.\textsuperscript{99}

Completeness of follow-up

Authors should always report the outcome of all patients who were entered into a study when their outcomes are known. When patients are lost to follow-up, their loss should be noted and explained and the impact of their absence on the statistical analysis of results should be determined and discussed.\textsuperscript{53, 45} If substantial numbers of patients are “lost to follow-up” and the impact of their omission on the outcome of the study is not accounted for, the validity of the study results is jeopardized.\textsuperscript{22} Patients who are lost to follow-up have been handled in many different ways in clinical trials.\textsuperscript{62} Dropouts may be analyzed in their respective groups in all feasible ways (i.e., they can be assumed to have good or bad outcomes and the impact of their inclusion on the results can be determined). Patients can be analyzed in the group to which they are originally assigned regardless of what happens to them. This choice leads to a so-called “decision-to-treat” analysis. Patients can be counted as an end result at the time of withdrawal (i.e., lost to follow-up) is an end point able to be analyzed for treatment and comparison groups). There are two additional ways in which dropouts are commonly handled, but they cannot be logically justified and they lead to biased or flawed results. Dropouts can be ignored or eliminated from the study and not counted in the end results. Because patients who drop out may have a significantly different prognosis and outcome than patients who complete a study, this treatment of dropouts may have a significant effect on the validity of the study.\textsuperscript{3} This problem is especially true when the numbers of dropouts in treatment and control groups are very different or when control and treatment group patients drop out for different reasons. Finally, in some studies, patients who drop out of one group may receive the alternative treatment and then are counted and analyzed in the group in which they end up. This practice can never be justified.\textsuperscript{62}

One way to determine whether the patients lost to follow-up had a significant impact on the validity of a clinical trial is to assume that all patients in the control group who were lost to follow-up improved and all the patients in the treatment group did poorly, and recalculate the results. If this exercise significantly altered the outcome and conclusion of the study, then the study’s validity may be questioned.\textsuperscript{22} This worst-case scenario gives the most negative impact that dropouts can have on a study’s result and may be overly conservative.\textsuperscript{33} Its utility in a particular study depends entirely on how likely the assumption (that all the control dropouts did well and all of the treated dropouts did poorly) approximates the truth.

The final four criteria deal with presentation and analysis of results. During medical school, most
discussions of medical statistics are met with dread or disinterest. The statistics required to understand and interpret clinical trials are the most simple and most fundamental. Ironically, these basic principles are the ones most commonly omitted or misused in published clinical trials. Physicians can understand the majority of statistics used in clinical trials by understanding the meaning of p value, knowing how to use the t test correctly, and understanding contingency tables.

**Reporting the data**

The authors of a study should present their results in sufficient detail to allow the reader to perform his or her own preferred analysis of the data. At a minimum, the presentation of interval data (i.e., discrete or continuous data such as lesion counts or area of involvement) should include a summary measure of the center of the data for each group and a measure of the variability of the data. The mean and median are most commonly used to indicate the center of the data. The mean is simply calculated as the sum of the data points divided by the total number of data points. The mean is the most appropriate choice for normally distributed data (i.e., data that are distributed in the shape of the bell curve). The median is a better choice for data that are skewed or contain outliers (Fig. 2, A and B). The median for an odd number of data points is the value in the middle when the data are ordered from lowest to highest. The median for an even number of data points is determined by taking the mean of the two numbers in the middle when the data points are ordered from lowest to highest. Half the data points fall above and half fall below the median. The magnitude of the difference between the mean and the median is a crude indicator of the degree of skewing. If the difference is none or small, the data are or approach normal (i.e., are symmetric); if the difference is large, the data are skewed.

The standard deviation (SD), range, interquartile range, and 95% confidence intervals are commonly used to indicate the variability or spread of the data (Fig. 2, B-F). The SD is the square root of the average squared deviation of values from the mean. That is:

\[ SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \]

SD is an important summary statistic for normally distributed data only. For normally distributed data, approximately 68% of the data points will lie within the range from the mean minus one SD to the mean plus one SD; 95% of the data points will lie within the range from the mean minus two times the SD to the mean plus two times the SD. These relations do not apply to skewed data or small data sets that contain outliers. Note in Fig. 2, C that the mean and SD do not adequately describe the center of the data or its distribution for skewed data or data with outliers (treatments Y and Z). The range is the interval from the highest to the lowest data points. The range may be a good indicator for variability for data that are tightly arranged around the mean and for small data sets. The range is greatly influenced by outliers and may not give adequate indication of skewed data. Therefore the range is not a good descriptive statistic for skewed data sets or for data with outliers (Fig. 2, A and B). The interquartile range is the interval that contains the middle 50% of the data points. It excludes the highest 25% of the data points and the lowest 25%. The interquartile range is a reasonable descriptive statistic to describe skewed data and data with outliers (Fig. 2, E). Confidence intervals are the best way to present the degree of uncertainty of normally distributed data as well as skewed data or data with outliers (see discussion of confidence intervals later). The 95% confidence interval around the mean can be calculated for normally distributed data with the sample mean, standard error, and the t distribution. For skewed data, the confidence interval around the median should be used (Fig. 2, F). The confidence interval provides a range of values in which the "population" or true response to treatment lies. For example, the 95% confidence interval is 95% likely to contain the population or true mean. If the 95% confidence intervals around two means or medians do not overlap, you can be confident at the 95% level that the two population means or medians, from which the samples were drawn, are different.

The standard error (SE) is the SD divided by the square root of the number of data points (SE = SD/\sqrt{n}). Unfortunately, the SE is commonly but incorrectly used to summarize the variability of data. This choice is made because of tradition and because the SE is always smaller than the SD and 95% confidence interval and, therefore, looks better (Fig. 2, C, and D, and F). The SE is correctly used to analyze normally distributed data in inferential statistics (see discussion of the t test lat-
Fig. 2. Hypothetical final lesion counts in clinical trial of treatments X, Y, and Z. The same data are displayed or summarized in each panel. A graphic display of all data points (A) gives an accurate description of the magnitude and distribution of the results. B, Commonly used summary statistics are given. The mean with error bars consisting of one standard deviation (C) or one standard error (D) provide little useful information even for the normally distributed response to treatment X and are totally misleading for treatment Y, which is skewed, and treatment Z, which has outliers. A box plot of the data (E) shows the median (line within box) and the interquartile range (top and bottom of box) and adequately represents skewed data (note asymmetry of median within box in Y plot) and presence of outliers (note asymmetry of median within box in the Z plot and the outliers, indicated by asterisk and open circle). The 95% confidence interval about the median (F) adequately represents skewed data and the presence of outliers. If the 95% confidence intervals around two means or medians do not overlap, you can be confident at about the 95% level that the two population means or medians are different.
ttributed data sets, there is a 95% chance that the true or population mean (from which the sample was drawn) will lie approximately within the range from the sample mean minus two times the SE to the sample mean plus two times the SE. The relation among the sample mean, the population mean, and the SE does not apply to skewed data or small data sets that contain outliers.69, 104, 107-110

Descriptive statistics are at best incomplete summaries of the data; considerably more information may be conveyed by graphic displays (Fig. 2).69, 107, 108 The production and interpretation of graphic displays are well reviewed by O'Brien and Shapiro.108 Figures that show individual observations are an excellent tool for presenting small data sets and for presenting data that are skewed or have outliers (Fig. 2, A).69, 107, 110 Unfortunately, such diagrams are underutilized in the scientific and medical literature. Graphic displays of data using “error bars” showing one SD or one SE depict at best 68% of the sample or 68% confidence intervals, respectively, and therefore are misleading and should be avoided or approached with skepticism (Fig. 2, C and D).69, 107

Categorical data are preferably presented in tables in which treatments and outcome are represented in rows and columns.106, 116 Examples of categorical outcomes include cured and not cured, survived and died, and minimal, moderate, and marked improvement. Categorical outcomes should be clearly defined, limited in number, clinically important, and declared before the trial is begun.

Statistical analysis

Observed differences in outcome in clinical trials may be caused by chance alone. It is therefore imperative that statistical analyses be performed to verify that the results are from treatment and not from chance. This admonition is especially true for trials containing few patients, as are many published dermatologic trials. This fact can be illustrated by a simple example using dice to simulate a clinical trial (Fig. 3). Suppose we simulate a clinical trial of a new therapy for metastatic melanoma with a new treatment (A) that has a cure rate of 33% and compare it to another treatment (B) that has the same cure rate. Treatment of a patient is simulated by throwing the dice. If it lands on 1 or 2 the patient lives; on 3, 4, 5, or 6, the patient dies. The results of 10 “trials” treating 10 patients with A and 10 patients with B are shown in Fig. 3, d. The results were obtained by means of Resampling Stats (a computer program that allows simple programs to be used to perform resampling procedures) and not by actually rolling dice.102 As demonstrated in Fig. 3, d, the results in some trials suggest that one treatment is superior to the other. However, by the nature of these “trials,” the differences are from chance alone. Statistical tests are employed to determine the probability that, given no treatment effect, a difference as large as the observed difference might arise by chance. Surprisingly, statistical analyses are often omitted from published clinical trials.19, 20, 104, 110, 117 The importance of sample size is also illustrated in Fig. 3, e-g. If 100 patients are treated with A and 100 with B in our simulated trial, the results of the two treatments are more nearly equal in all 10 “trials.”

Statistical methods

Understanding p

Understanding the meaning of p is fundamental to understanding and interpreting statistical testing.104, 118 In simplest terms, all clinical trials begin with the null hypothesis that the treatments are the same. If the results of the trial indicate that the outcomes of treatment are different, then a statistical test is performed to determine whether the difference can be from chance or sampling variation alone. If chance is not a likely explanation of the difference, the null hypothesis that the treatments are the same is rejected.33, 118, 119 The p value represents the probability that a difference as large as the observed difference might arise by chance if the treatments are the same. The error of believing that a difference in treatments exists, when in fact there is no difference in treatments and the observed differences were from chance or sampling variation alone, is referred to as a type I or α error. The statement that “the difference was statistically significant (p = 0.05) means that the probability that a difference as large as the observed difference might arise by chance is 1 in 20 and that the author will accept this degree of chance as being sufficiently unlikely that the null hypothesis that the treatments are the same can be rejected. Conversely, the statement that “the difference was not statistically significant (p = 0.10) means that the probability that a difference as large as the observed difference might arise by chance was 1 in 10 and that the author will not accept this degree of chance as being sufficiently unlikely that the null hypothesis that the treatments are the same can be rejected. It is important to remember that not significant means that a
Fig. 3. Resampling Stats program using dice to simulate a clinical trial. A clinical trial of treatment for metastatic melanoma with a new treatment (A) and standard treatment (B), both of which have cure rates of 2/6, is simulated by throwing dice. If the die lands on 1 or 2, the patient lives; on 3, 4, 5, or 6, the patient dies (a). The program that simulates this trial is shown in (b) and the explanation for each line in (c). Numbers in the tables (d and g) represent the number of patients who survive. Each column represents a separate trial. Some trials with 10 patients in each group (d) appear to demonstrate that one treatment is superior to the other (asterisk). When 100 patients are studied (e-g), the results of both treatments are more likely to be similar (g).

The acceptance of a significance level of 0.05 as the cutoff for rejecting the null hypothesis is a tradition based on quality control standards and is not an absolute truth. At times more stringent standards are used in medical research. When the results are close to the boundary of the region (p ≤ 0.05), it is sometimes worthwhile to collect more data, since a small difference is not proved; it does not mean that there is no difference.

Meet the p = 0.05 standard sometimes may be significant. For example, a trial of a new chemotherapeutic agent involving 30 randomized patients with metastatic melanoma produced a 5-year survival rate of 47% in patients treated with the new agent (7 of 15) and 20% in control patients treated with conventional non-chemotherapy and radiotherapy.
ation (3 of 15). Whereas the result does not achieve statistical significance when compared by chi-square testing (see later) (Yates-corrected \( \chi^2 = 1.35; p = 0.25 \)), the result is, nonetheless, potentially significant. If the therapy is beneficial and the estimated difference in response rates is the true difference in response rates, it may result in the saving of 2400 lives annually (based on 7200 deaths from melanoma annually and the improvement in survival in this hypothetical example). Because of the biologic and clinical importance of the results suggested by the study, the treatment should be subjected to study in a larger patient population with more power to detect a difference if one exists (see discussion of "Power"). Others might argue that the adherence to the 95% standard (\( p = 0.05 \)) should be relaxed in cases like this one. The potential benefit of the treatment may be revealed by the use of confidence intervals (see discussion of "Confidence Intervals").

Investigators should indicate the statistical procedures that were utilized to compare results. Simply stating that the difference was statistically significant (\( p < 0.050 \)) does not constitute an adequate description of the statistics used. The exact procedure used (e.g., \( t \) test or \( \chi^2 \)) and the results obtained must be specified. The \( t \) test and the analysis of contingency tables (e.g., chi-square and Fisher exact tests) are the most commonly employed procedures and should be understood by all physicians.

The \( t \) test

The \( t \) test is the most commonly used statistical test in the biomedical literature. The \( t \) test is used to determine whether the difference in the means of two samples is from chance (sampling variation) alone. It bears emphasis that the \( t \) test is designed to compare the difference in the means of only two samples or populations. The \( t \) test has additional important restrictions. The \( t \) test should be used to compare the means of interval data that are normally distributed. The data analyzed by the \( t \) test should be sufficiently large, should not contain outliers, and should be statistically independent; in addition, the variability of the two treatment groups should be similar. In most instances a two-sided \( t \) test is appropriate. When a one-sided \( t \) test is performed, the reasons should be specified.

The unpaired \( t \) test should be used to analyze two different groups (e.g., patients receiving alternative treatments). The paired \( t \) test should be used to analyze the effects of two different treatments on the same patients (e.g., in a crossover or left-right comparison study). Misuse of the \( t \) test was one of the most commonly encountered errors in the use of statistics found by Gore, Jones, and Ryttner in their review of the use of statistical methods in the British Medical Journal. They found 52% of analytical reports (32 of 62) that utilized statistical analyses contained at least one error in the use of statistics. Eleven of the errors were misuse of the \( t \) test. The errors included using the \( t \) test to analyze data that were not normally distributed or that had different variability or using an unpaired \( t \) test to analyze paired data. Another commonly encountered error in the use of the \( t \) test is its use to compare the means of more than two samples. This error in the use of the \( t \) test occurred in 61% and 44% of articles that used statistical analyses published in Circulation Research and Circulation, respectively, in 1977.

Statistical analysis with the \( t \) test can be illustrated by using data from a study comparing calcipotriol and betamethasone ointments in the treatment of stable plaque-type psoriasis. In this multicenter randomized control trial, 201 and 200 patients were randomly assigned to treatment with calcipotriol and betamethasone, respectively. Outcomes were measured by changes in PASI scores. At the end of 6 weeks of treatment the improvement (decrease) in PASI scores was 5.50 and 5.32 for calcipotriol and betamethasone, respectively. The statistical significance of the small difference in the change in PASI scores was evaluated by means of an unpaired \( t \) test. The \( t \) value and corresponding significance level can be determined with computer programs or calculated with a \( t \) distribution table and the formula. Not surprisingly, the result indicated that the difference was not statistically significant (\( p = 0.76 \)). Thus the trial failed to demonstrate a difference between the two treatments.

The \( t \) statistic is:

\[
t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}
\]

in which \( \bar{X}_1 - \bar{X}_2 \) is the difference between means, \( n_1 \) and \( n_2 \) are the numbers of patients in each treatment group, and \( S^2 \) is the pooled estimate of the variance (SD squared) given by the equation:

\[
S^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}
\]

in which \( SD_1^2 \) and \( SD_2^2 \) are the SDs of each treatment group squared.
Table II. Format of a fourfold contingency table

<table>
<thead>
<tr>
<th>Treatment A</th>
<th>Good response</th>
<th>No or poor response</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>H1</td>
<td></td>
</tr>
<tr>
<td>Treatment B</td>
<td>c</td>
<td>d</td>
<td>H2</td>
</tr>
<tr>
<td>Total</td>
<td>V1</td>
<td>V2</td>
<td>N</td>
</tr>
</tbody>
</table>

The chi-square for unpaired data is calculated using the formulas:

\[
\chi^2 (\text{uncorrected}) = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{(a + b)(c + d)(a + c)(b + d)}
\]

\[
\chi^2 (\text{Yates corrected}) = \frac{(a \cdot d - b \cdot c - 0.5 \cdot N)^2 \cdot N}{(a + b)(c + d)(a + c)(b + d)}
\]

in which \((a \cdot d - b \cdot c)\) indicates the absolute value of \(a\cdot d - b\cdot c\).


The \(t\) test should not be used to analyze data that are skewed or data in which the variability (SD) is very different between treatment groups.\(^{107,113}\) Nor should it be used to analyze small data sets because it is impossible to verify that small data sets are normally distributed.\(^{113}\) In these situations, the Mann-Whitney U test, the distribution-free equivalent of the \(t\) test, may be used for unpaired data and the Wilcoxon test can be used for paired data.\(^{109,112,114,128}\) The \(t\) test also should not be used to compare more than two groups because it ignores possible associations among groups and it increases the likelihood that spurious results will be obtained.\(^{104,113,119,126}\) Testing of means of more than two groups can be performed with different forms of analysis of variance.\(^{104,107,113,126}\)

Contingency tables

The chi-square test is the most commonly used technique to determine whether the categorical data in contingency tables differ because of chance alone.\(^{106}\) The format of a typical fourfold (2 × 2) contingency table is shown in Table II. The numbers in the table (indicated by the letters a, b, c, and d) are the numbers of patients in each treatment group that fall into each outcome category. The assumption made to formulate the null hypothesis that the treatments are the same is that the overall response rate (i.e., the total number of patients with good responses divided by the total number of patients; \(V1/N\) in Table II) is the same for both treatments. Chi-square is then used to determine the likelihood that the actual numbers observed could be obtained by chance or treatments were the same.\(^{109,112,114,128}\) A continuity correction in the chi-square is recommended to analyze smaller data sets (20 to 40 patients).\(^{112,128}\) This correction is used to account for the use of continuous curves to approximate the discrete frequencies in contingency tables.\(^{112,128}\) The Yates correction is most commonly used (Table II). The chi-square test should not be used to analyze small trials (trials with fewer than 20 patients) or when the predicted number of patients in any cell (positions a, b, c, or d in Table II) is less than 5.\(^{3,112,114,116}\) In those circumstances, the Fisher exact test should be performed.\(^{109,112,114,128}\) The chi-square test is used to analyze unpaired data and McNemar’s test is used to analyze paired contingency table data.\(^{109,112,114,128}\) Contingency table analyses can be easily modified and used to compare more than two treatments and more than two outcomes.\(^{109,112,114,128}\)

The statistical analysis of contingency table data can be illustrated using data from a study comparing ketoconazole shampoo to placebo in the treatment of seborrheic dermatitis (see example in Fig. 4).\(^{129}\) In this study 36 randomly allocated patients with seborrheic dermatitis were treated with ketoconazole shampoo (18 patients) or placebo (18 patients). The results indicate that 14 of 18 patients treated with ketoconazole were cured, whereas only 2 of 18 patients treated with placebo were cured. The null hypothesis would assume that the overall cure rate (44% [16 of 36]) was the same for both treated and control groups and would predict that 8 of 18 patients in the ketoconazole group and 8 of 18 patients in the placebo group would be cured. Therefore the predicted values for the fourfold contingency table would be 8, 10, 8, and 10 for a, b, c, and d, respectively (see Table II). The statistical question to be answered is that if the treatments were the same and the overall cure rate is 16 of 36 patients, how likely is it that chance alone is responsible for the observed results? The result can be obtained by looking up the chi-square value in tables, calculating the chi-square by hand or with a calculator, or by calculating the chi-square with statistical software.\(^{116,130}\) The \(p\) value corresponding to the calculated \(\chi^2\) can be obtained from readily available tables or will be provided if the calculation is performed with statistical software.\(^{116,130}\) The Yates-corrected \(\chi^2\) is 13.6 and \(p < 0.001\). Thus the probability that a difference as large as the observed difference might arise by chance if there is no difference in treatments is less than 1 in 1000, and this result is sufficiently
a. Trial Results

<table>
<thead>
<tr>
<th>KETOCONAZOLE</th>
<th>Cured</th>
<th>Not cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>14</td>
<td>4</td>
</tr>
</tbody>
</table>

b. Statistical Result

Corrected chi-square = 13.6

P < .001

c. Program

d. Explanation

Repeat 1000
generate 18 \( a \times 36 \) a
count \( a < 17 \) c
count \( b < 17 \) d
divide \( c \times 18 \) cc
divide \( d \times 18 \) dd
subtract \( c \) cc from \( d \) dd e
score \( e \) f

diffit = as a cure.
determine the cure rate
for group \( a \) and \( b \).
determine the difference in rates.
keep track of the difference in rates.
repeat until 1000 trials
are performed.

'How often is the difference \( \geq .665 \)'
'Calculate \( P \).
'Print \( P \).
'draw a histogram of the difference in rates.

e. Results

\( P = 0 \)

<table>
<thead>
<tr>
<th>Diff in ( RR )</th>
<th>Freq</th>
<th>Pct</th>
<th>Cum Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>-.6</td>
<td>2</td>
<td>.2</td>
<td>.2</td>
</tr>
<tr>
<td>-.5</td>
<td>1</td>
<td>.1</td>
<td>.3</td>
</tr>
<tr>
<td>-.4</td>
<td>7</td>
<td>.7</td>
<td>1.0</td>
</tr>
<tr>
<td>-.3</td>
<td>51</td>
<td>5.1</td>
<td>6.1</td>
</tr>
<tr>
<td>-.2</td>
<td>138</td>
<td>13.8</td>
<td>19.9</td>
</tr>
<tr>
<td>-.1</td>
<td>237</td>
<td>23.7</td>
<td>43.6</td>
</tr>
<tr>
<td>0</td>
<td>147</td>
<td>14.7</td>
<td>58.3</td>
</tr>
<tr>
<td>.1</td>
<td>216</td>
<td>21.6</td>
<td>79.9</td>
</tr>
<tr>
<td>.2</td>
<td>134</td>
<td>13.4</td>
<td>93.3</td>
</tr>
<tr>
<td>.3</td>
<td>53</td>
<td>5.3</td>
<td>98.6</td>
</tr>
<tr>
<td>.4</td>
<td>11</td>
<td>1.1</td>
<td>99.7</td>
</tr>
<tr>
<td>.5</td>
<td>2</td>
<td>.2</td>
<td>99.9</td>
</tr>
<tr>
<td>.6</td>
<td>1</td>
<td>.1</td>
<td>100.0</td>
</tr>
</tbody>
</table>

f. Histogram

![Histogram](image)

Fig. 4. Results of a randomized clinical trial of ketoconazole shampoo versus placebo in treatment of seborrheic dermatitis (a),^12^ analyzed with hypothesis testing (b), and simulated by resampling (c-f). Note that the outcome achieved in this trial was never produced in 1000 resampled trials (arrows in f). Asterisk. Difference in response rates.

unlikely that the null hypothesis that the treatments are the same can easily be rejected.

The same problem can be calculated and understood by resampling. In this case, the question being asked is: "What is the probability that two different treatments that have response rates of 16 of 36 will produce a difference in response rates equal to or greater than the results of this trial?" Treatment of a patient is simulated by throwing a 36-headed die. If it lands on 1 through 16 the patient is cured; on 17 through 36, the patient is not cured. Each trial is simulated by throwing the die 18 times for ketoconazole and 18 times for the placebo. If this simulated trial is repeated 1000 times, how often will the cure rate of ketoconazole (78%) exceed the cure rate of placebo (11%) by 67% or more (or how often will the cure rate of placebo exceed the cure rate of ketoconazole by 67% or more)? The Resampling Stats program that simulates this trial is shown in Fig. 4, c and the results are shown in Fig. 4, e and in graphic form in Fig. 4, f. The results indicate that in 1000 simulations, a difference as great as the observed
difference did not occur. Therefore the probability that a difference as large as the observed difference might arise by chance is less than 1 in 1000. This result is sufficiently unlikely that the null hypothesis that the treatments are the same can easily be rejected.

Like the t test, the chi-square test is frequently used incorrectly in published clinical trials. In their analysis of the use of statistics in medical papers published in the British Medical Journal, Gore, Jones, and Rytter found that chi-square was used incorrectly in 12 of 62 published papers. The errors included omission of the continuity correction for studies with few patients, omission of a clearly stated hypothesis to be tested, lack of consideration of degrees of freedom, and misuse of the test to study paired data.110

**Confidence intervals**

Standard statistical hypothesis testing generally results in a determination of a p value, which is the probability that a difference as large as the observed difference might arise by chance. If this probability is equal to or less than a predetermined "critical value" (usually 0.05), the null hypothesis is rejected and the difference is considered statistically significant. If the p value exceeds the predetermined value, the null hypothesis is not rejected and the difference is considered not statistically significant. This dichotomous treatment of the results of clinical trials as being significant or not significant may sometimes be misleading and may hide meaningful estimates in the differences between the treatments.14, 44, 113, 116, 118, 132, 133 What matters most in a therapeutic trial is whether investigators have been able to detect a medically significant difference in treatments and how large the difference is likely to be.113, 132, 133 The conventional significance levels are useful guides to the interpretation of trial results, but they should not be considered strict rules.44 The absurdity of the dichotomous significance testing approach is best appreciated by considering two trials, one of which has a p value of 0.05 and the other a p value of 0.06. The difference in the level of significance is very small, yet the former is considered statistically significant and the latter is not.

Reporting results with confidence intervals is an alternative or complementary way to present the results of clinical trials. Many believe it is far preferable.133, 134 In simple terms, the reported result provides the best estimate of the treatment effect, and the confidence interval provides a range of values in which the "population" or true response to treatment lies.21, 22, 113, 135 For example, the 95% confidence interval is 95% likely to contain the population or true mean. Alternatively, if the trial is repeated many times, 95% of the confidence intervals produced will contain the true or population mean response to treatment. The true response rate will most likely lie near the middle of the confidence interval and will rarely be found at or near the ends of the interval. It will be beyond the limits of the 95% confidence interval only 5% of the time.21

The confidence interval provides a direct, numeric measurement of the imprecision of the estimate of the response to treatment that is due to sampling variability.132 The width of the confidence interval is determined by the sample size (the larger the sample, the narrower the interval), the variability of the response being measured (the larger the variability, the wider the interval) and the degree of confidence desired (the higher the confidence desired, the wider the interval) (see Fig. 5).21, 22, 113, 132 Confidence intervals can be determined for means and their differences, and for proportions and their differences by looking them up in tables, by calculating them with formulas, or by using computer software.3, 21, 111-114, 116, 136, 137 In comparative studies, confidence intervals should be reported for the differences between groups, not for the results of each group separately.107, 114

There is a close relationship between the results of a test of a hypothesis and the associated confidence interval; if the difference between treatments is significant at the 5% level, then the associated 95% confidence interval excludes the zero difference.21, 107, 113, 119 The advantage of using confidence intervals instead of or in addition to p values is that confidence intervals provide an indication of the size of the differences in treatments and give numeric measurements of the inexactness in our knowledge of the real differences in treatments.113

To illustrate the utility of confidence intervals in differences in means, let us return to the comparison trial of calcipotriol and betamethasone cited earlier.137 In this study, the difference in the mean change in PASI scores for calcipotriol and betamethasone was 0.18 (5.50 - 5.32) and the 95% confidence interval of the difference was -0.40 to 0.78. Because the 95% confidence interval includes the zero difference, we cannot be 95% certain that the effects of calcipotriol and betamethasone are different. The difference is most likely to be 0.18 but may be as low as -0.40 or as high as 0.78.
Fig. 5. Confidence interval is dependent on sample size, sample variability, and confidence level desired. In the top panel, the mean change in PASI score was 3.1 with an SD of 2.9. The 95% confidence interval is shown and narrows as the number of patients studied increases from 20 to 200. In the middle panel, the mean change in PASI score of 20 patients was 3.1. The 95% confidence interval is shown and widens as the sample variability (SD) increases. In the lower panel, the mean change in PASI score of 20 patients was 3.1 with an SD of 2.9. The confidence interval widens as the degree of confidence desired increases from 90% to 99%.

as –0.40 (favoring betamethasone) or as high as 0.78 (favoring calcipotriol).

To illustrate the utility of confidence intervals in differences in proportions, the confidence interval for the difference between placebo and ketoconazole shampoo for the treatment of seborrheic dermatitis cited earlier can be calculated and interpreted. In this example the cure rate for ketoconazole was 78%, the cure rate of placebo was 11%, and the difference between cure rates was 67%. The 95% confidence interval for the difference in proportions is 0.27 to 0.82. Because the 95% confidence interval of the difference excludes the zero difference, we can be 95% certain that the response rates of ketoconazole...
Fig. 6. The resampling program (a), explanation (b), results (c), and histogram (d) for determining the 95% confidence interval of the results of the ketoconazole shampoo versus placebo trial. The 95% confidence interval (CI) is indicated by arrows in the results and the histogram. Asterisk: Difference in response rates.

in treatments is illustrated by the fact that this study has been cited more than 700 times. Despite the recognition of the importance of power, a follow-up study conducted by Moher, Dulberg, and Wells in 1994 indicated that 25% or 50% improvements in outcome might have been missed in 84% and 64% of 102 negative studies, respectively.

Power is an expression of the ability of a trial to detect a difference in treatments if one exists. The investigators in a well-designed clinical trial should have an estimate or prediction of the anticipated differences among treatments and knowledge of the magnitude of differences that would be clinically important. The trial should then be designed to enroll a sufficient number of patients to ensure with reasonable certainty that a statistically significant difference would be obtained if the anticipated or clinically important differences between the treatments existed. The error of believing that there is no difference between treatments, when in fact there is a difference, is referred to as a type II or β error.

The power of a trial to detect a significant difference is determined by the response rates of the treat-
ments, the significance level desired by the investigators, and the number of patients treated.\textsuperscript{109,141} The power of a trial to detect differences between treatments can be calculated with formulas, nomograms, programmable calculators, or computer software.\textsuperscript{13,143-146} Conventionally, a power of 80\% with a significance level of 0.05 is considered adequate.\textsuperscript{116,119} The choice of a power of 80\% means that, if the anticipated or clinically important difference in treatments really exists, four of five trials with the specified number of patients in the treatment groups will show a statistically significant difference. However, higher power sometimes may be desirable. The power of a study design to detect clinically significant differences in treatments is best considered before the trial is initiated.\textsuperscript{141,142} However, studies that fail to demonstrate statistically significant differences among treatments should always include a discussion of power.

The meaning of power can be illustrated by using another resampling example that simulates treating patients by throwing dice. Suppose that in our simulated clinical trial of a new treatment of metastatic melanoma, the old treatment (A) has a cure rate of 33\% (two of six patients) and the new treatment (B) has a cure rate of 67\% (four of six patients). Treatment of a patient is again simulated by throwing the dice (Fig. 8). If the die simulating treatment A lands on 1 or 2, the patient lives; on 3, 4, 5, or 6, the patient dies. If the die simulating treatment B lands on 1, 2, 3, or 4, the patient lives; on 5 or 6, the patient dies. The results of 10 "trials" treating 10 patients
Table III. Sample size required in each of two treatment groups*

<table>
<thead>
<tr>
<th>Smaller of two percentages expected to respond</th>
<th>Larger of two percentages expected to respond</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>20</td>
<td>332</td>
</tr>
<tr>
<td>30</td>
<td>395</td>
</tr>
<tr>
<td>40</td>
<td>426</td>
</tr>
<tr>
<td>50</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
</tr>
</tbody>
</table>


*To ensure an 80% probability of finding a statistically significant difference when \( p = 0.05 \) for sample sizes up to 100 in a two-tailed test.
†Sample size needed to detect a difference is small when the difference in treatments is large.
‡Sample size needed to detect a difference is large when the difference in treatments is small.
§Sample size needed to detect a difference is large when the response rate of the comparison treatment is high.

with A and 10 patients with B are shown in Fig. 8, d. As demonstrated in Fig. 8, d, the results in most trials suggest that treatment A is not significantly different from treatment B. If 100 patients are treated with A and 100 with B, the difference in the two treatments are more clearly demonstrated (Fig. 8, e-g). The power of the trial with 10 patients in each group is low (i.e., the significant difference in the response rates is likely to be missed in the trial because the number of patients studied is inadequate). In contrast, the power of the trial with 100 patients is very high (i.e., it is very unlikely that the difference in response rates would be missed in any trial enrolling this number of patients). Specifically, the power of the 10-patient trial is only 17% (i.e., only 17 of 100 trials enrolling 10 patients per group would detect a statistically significant difference); and the power of the 100-patient trial is 99.8% (i.e., 998 out of 1000 trials enrolling 100 patients would detect a statistically significant difference). An investigator who wanted the power of his or her study to detect a difference in treatments of this magnitude to be 80% with a significance level of 0.05 would have to enroll 38 patients in each group.†

The meaning of power can also be illustrated by a sample size table (Table III).‡ As indicated by the single dagger, the sample size needed to detect a difference is small when the difference in treatments is large. Conversely, the sample size needed to detect differences is large when the difference in treatments is small or when the response rate of the comparison treatment is high (points ‡ and § respectively).

As noted by Moher, Dulberg, and Wells, 142 “If a trial with negative results has a sufficient sample size to detect a clinically important effect, then the negative results are interpretable—the treatment did not have an effect at least as large as the effect considered to be clinically relevant. If a trial with negative results has insufficient power, a clinically important but statistically insignificant effect is usually ignored or worse, is taken to mean that the treatment under study made no difference. Thus, there are important scientific reasons to report sample size and/or power calculations.” (p. 123)

Subgroup analysis

Authors of studies whose primary hypothesis produces negative results may analyze subgroups of patients in an attempt to find a clinically important and statistically significant result.21,119 These subgroup analyses may have methodologic problems including the use of multiple comparisons with no adjustment in required significance levels, treatment of dependent variables as though they were independent, and overinterpretation of data.21,33,118,119 The features of subgroup analyses that make them most likely to be valid include observation of large effects, results that are unlikely to be due to chance, results studied because the analysis was hypothesized before the study began, one of only a few subgroup analyses, and results that are replicated in other studies.21,33,148 Subgroup analyses may be useful to generate hypotheses to test in future studies; they

*These power and sample size calculations were performed with STPLAN, a computer program written by Barry W. Browne and colleagues to perform power, sample size, and related calculations needed to plan studies. The program is available on the Internet free of charge by anonymous FTP from odin.mda.uth.tmc.edu.
should not, however, be relied on as sources for determining the best available treatment.

The importance of statistical analyses must be kept in proper perspective. Statistics are a tool for trying to ensure that results of clinical trials are not due to chance or sampling variation alone. The combination of hypothesis testing and the use of confidence intervals give a measure of the likelihood that results of a trial are due to chance and the precision of the estimated difference in treatments, respectively. Statistical analyses cannot tell you the medical significance of differences in treatments. In other words, "statistically significant" should not and cannot be equated with "medically significa-
Table IV. Shortcut method

<table>
<thead>
<tr>
<th>Step</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read title</td>
<td>If interesting, continue.</td>
</tr>
<tr>
<td>Read abstract</td>
<td>If still worthy of your time, continue.</td>
</tr>
<tr>
<td>Look at figures and tables</td>
<td>Find most important data. Are the data adequately reported? Is there a control group? Are outcomes clinically important? Are data subjected to statistical testing and are tests used correctly? If still worthy of your time, continue.</td>
</tr>
<tr>
<td>Read results</td>
<td>Are complications reported? Are compliance and patients lost to follow-up considered? Ask yourself: “If everything the authors say is true, is it important for my patients, my practice, or for the field of dermatology?” If “no,” stop reading; if “yes” or “possibly,” continue.</td>
</tr>
<tr>
<td>Read methods</td>
<td>How were patients chosen and allocated? Are outcomes clearly defined and assessed blindly? Are the conclusions valid? If yes, continue.</td>
</tr>
<tr>
<td>Record citation</td>
<td>Note authors, citation, and conclusions of the study for easy retrieval later.</td>
</tr>
<tr>
<td>Read discussion</td>
<td></td>
</tr>
<tr>
<td>Read introduction</td>
<td></td>
</tr>
</tbody>
</table>

cant,"14, 69, 107, 116, 119 Conversely, a trial whose results indicate that the difference in treatments was not statistically significant does not necessarily mean that there is not a medically important difference between treatments. It may simply mean that too few patients were used to detect the difference that does exist.116,141

SHORT-CUT METHOD FOR READING CLINICAL TRIALS

A short-cut method to evaluate clinical trials will enable you to use your time efficiently (Table IV). It allows you to decide not to read the majority of poorly conceived, designed, executed, or reported trials and those trials with insignificant results. After each step you should reask the question, “Am I still interested in this trial?” and “Is it worth spending my valuable time on it?” You should be looking for the 15 items of methodology listed in Table I to see whether they were reported and used properly.14

First read the title. If the title holds interest for you, read the abstract. If after reading the abstract you still think that the article is worthy of your time, look at the figures and tables. Authors generally display their best or most convincing data in figures and tables.69 In a well-written study, you should be able to see the important results in summary figures and tables. Try to pick out the one or two figures or tables that hold the most important data. Look at the data to see whether it is summarized and displayed correctly.14 Determine whether there is an appropriate control group, whether the outcomes are clinically and biologically important, whether the data are subjected to statistical analysis, and whether the correct statistical methods were used. It is important that you examine and assess the data yourself.14 If you are still convinced that the article is worthy of your time, read the results. Determine whether complications are reported, and whether compliance and lost patients were considered. After reading the results ask yourself the question, “If everything the authors say is true, is it important for my patients, for my practice, or for dermatology as a whole?” If the answer is an emphatic “no,” stop reading; if “yes or possibly,” continue to the methods.

While reading the methods pay particular attention to the features that strengthen clinical trials and help validate their conclusions. In particular, determine the eligibility criteria, whether patients were randomly allocated, and whether the outcome is clearly defined and assessed blindly. If your conclusion is that the trial was conducted in a reasonable manner and the conclusions are valid, then make some notation of the authors, the source, and the conclusions of the study. Several software packages for cataloguing and managing references are available for most computer platforms (e.g., EndNote and Reference Manager). Finally, if you still have time
and interest, read the discussion and introduction, in that order.

OTHER METHODS FOR ASSESSING CLINICAL TRIALS

The method for reviewing clinical trials that we have described is comprehensive and can help the reader understand and interpret clinical trials. Its utility in evaluating trials in the medical and dermatologic literature has been validated. However, the method has some disadvantages. It leaves the reader to determine the overall quality of the trial, whether the methodology employed and reported is adequate, and whether the trial's conclusions are valid without providing specific guidelines for making these decisions. The criteria used are not weighted to give the user a sense of the relative importance of the 15 criteria. Many criteria are considered and using them to evaluate clinical trials requires some knowledge, understanding, and experience. Finally, the method does not result in a numeric score that can be used to grade the overall quality of the trial or the validity of the results. For these reasons, we briefly review other methods for assessing clinical trials (see the annotated bibliography after the references). Many of these methods overcome some of the deficiencies of the method we have chosen to present. Readers interested in utilizing any of these other methods are encouraged to review the original description of the method, referenced manuscripts where the methods have been used, and Spilker's excellent review of systems used to evaluate published data.

PROBLEMS WITH OTHER SOURCES FOR TREATMENT RECOMMENDATIONS

Pharmaceutical companies may have results of properly conducted clinical trials that they have chosen not to publish. The results of these unpublished trials may have been used to satisfy FDA requirements for providing the safety and efficacy of therapy. Although the data in these trials may be as good or better than trials published in peer-reviewed journals, special effort is required to obtain them and often they are unavailable for general public examination.

Advertisements in medical journals are potential sources for treatment recommendations, and physicians often obtain information by reading them. However, they have limited value because they often contain misleading data, figures, text, and graphics. In a systematic evaluation of 109 pharmaceutical advertisements in medical journals, Wilkes, Doblin, and Shapiro found that pharmaceutical advertisements commonly misidentified the advertised drug as being "superior" or the "drug of choice," lacked balance, had inadequate citing of relevant studies, and had misleading headlines and graphs.

Ziegler, Lew, and Singer systematically evaluated the accuracy of drug information provided by pharmaceutical sales representatives at medical teaching conferences using stringent criteria for inaccuracy. They found that 11% of statements made by the representatives about their drugs were inaccurate. However, most representatives will provide you with copies of published clinical trials if you ask them. However, the selection of trials obtained through drug company representatives may be heavily biased in favor of the representatives' products.

Sponsored medical symposia are commonly published in many peer-reviewed journals and contain extensive amounts of information about single therapeutic agents. In an evaluation of sponsored symposia published in peer-reviewed journals, Bero, Galbraith, and Rennie found that symposia were generally sponsored by a single pharmaceutical company, were less likely to be peer reviewed, used brand names of drugs, had misleading titles, and promoted unapproved indications. The authors concluded that using published symposia as sources of treatment recommendations should also be approached with skepticism. Rochon et al. expressed similar concerns about articles published in journal supplements. Their evaluation indicated that articles published in supplements were generally inferior. Supplement articles were also more likely to be sponsored by the pharmaceutical company and to contain references from other supplements. Concerns about the quality of articles published in supplements have been expressed by others. Pharmaceutical company-sponsored continuing medical education courses are commonly given and are a potential source for treatment recommendations. However, these courses may contain biases favoring the sponsoring company's drugs and may adversely affect physician prescribing patterns.

Relying on experts' opinions and consensus-
derived standards of care may not provide the best evidence currently available.\textsuperscript{4,163} For example, Antman et al.\textsuperscript{163} conducted a study comparing meta-analysis of treatments of acute myocardial infarction and recommendations in review articles about the acute treatment and secondary prevention of myocardial infarction. They found that in five of six instances in which published data revealed the efficacy of treatments in reducing mortality in acute myocardial infarction, there was a delay of several years before experts consistently recommended the therapies. Similar delays were documented for secondary prevention.

Physicians rely on their personal experiences to determine ways to treat patients and which therapies to use, but their determinations are subject to several pitfalls.\textsuperscript{13} One is selective recall.\textsuperscript{13} Physicians may remember patients who improved, often assume that patients who did not return for follow-up improved, and conveniently forget the patients who did not improve. Another pitfall is "avoiding" the last disaster.\textsuperscript{13} For example, a physician may treat a patient with acne with minocycline and the patient may have a severe life-threatening hypersensitivity reaction. This physician may avoid using minocycline for many future patients with acne, although minocycline may be more efficacious and less toxic than the alternative treatment that the physician chooses. A third pitfall is that few physicians keep adequate, easily retrievable records to codify results of treatments with a particular agent or of a particular disease; even fewer actually carry out analyses. Few physicians make provisions for tracking those patients who are lost to follow-up. Finally, for many conditions, a single physician sees far too few patients to enable reasonably firm conclusions to be drawn about the response to treatments.\textsuperscript{3} For example, suppose a physician has treated 20 patients with lichen planus with tretinoin and found that 12 (60\%) had an excellent response. The confidence interval for this response rate (i.e., the true response rate for this treatment in the larger population from which this physician's sample was obtained) ranges from 0.36 to 0.81. Thus the true response rate might well be substantially less (or more) than the physician concludes from personal experience.

**ROLE AND LIMITATIONS OF THE FDA**

The FDA is charged by law with the responsibility of ensuring that drugs and biologic agents that are approved for marketing in the United States are effective for their labeled indications, provide benefits that outweigh their risks, are of high quality, and have directions for use that are complete and honestly communicated.\textsuperscript{164-167} Achieving these goals is the responsibility of the Center for Drug Evaluation and Research where new drug applications are reviewed by panels of physicians who determine whether submitted clinical studies are adequate and well controlled and whether the results demonstrate that the drug is safe and effective for its proposed indication. Studies are also reviewed by statisticians to evaluate the studies' designs, the validity of the statistical analyses, and the conclusions of safety and effectiveness based on the study data.\textsuperscript{165} The FDA believes that showing of effectiveness in well-controlled clinical trials must be replicated to constitute an adequate demonstration of effectiveness for a new product. However, on occasion, the FDA has approved drugs on the basis of a single, well-designed, multicenter study.\textsuperscript{166}

Physicians can be assured that high-quality evidence of a drug's safety and efficacy has been provided to gain FDA approval of a drug. However, the FDA approval process has several inherent limitations that make it unwise for physicians to abdicate responsibility for evaluating a drug's use, safety, and effectiveness. The FDA has been under tremendous pressure by industry and some administrations to make the process of introducing new drugs, devices, and biologic agents less burdensome in terms of cost and time.\textsuperscript{166} As a result, the FDA has allowed efficacy to be established using surrogate end points because statistically and clinically significant differences can be demonstrated in smaller studies with surrogate end points. Drugs approved on the basis of surrogate end points may not prove to be effective when clinical end points are used. The FDA has also allowed efficacy to be established using placebo controls because it is much harder to demonstrate differences between two active agents and much larger studies and tighter designs are required to try to prove equivalence. Drugs approved on the basis of placebo controls may not prove to be as or more effective than existing therapies.

Drugs may win approval for a narrow and specific indication or even for use in specific subsets of patients. Once approved, however, the drug may be promoted and used for wider indications and larger populations of patients.\textsuperscript{166,168} For example, itracon-
azole was approved by the FDA solely for the treatment of systemic mycoses, but once released, it was advocated and used for the treatment of a wide variety of superficial fungal infections. As another example, topical tretinoin was approved by the FDA in 1972 for the treatment of acne but it is now widely used to ‘‘treat’’ wrinkles (an unapproved indication) and to prevent photoaging and skin cancer. Under these circumstances, the responsibility for determining the effectiveness and safety of drugs and the ways in which they will be used lies with physicians and not with the FDA.

CONCLUSIONS

Ultimately, it is the responsibility of the prescribing physician to determine the indications, safety, and efficacy of the drugs he or she uses. Well-designed, controlled clinical trials are the best source for determining the best treatments. We have tried to help dermatologists find and use the tools necessary to understand and evaluate clinical trials.

We thank Paul Bleicher, MD, PhD, Peter Bruce, PhD, and Gunnar Nyberg, MD for reviewing the manuscript and Alfred Allen, MD, PhD, Gunnar Nyberg, MD, Mary Evans, MD, Robert Haynes, MD, PhD, Stanton Giantz, PhD, and Thierry Poynard, MD for permission to use their ideas, figures, or tables.

REFERENCES

21. Guyatt GH, Sackett DL, Cook DJ. Users’ guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994;271:59-63.
22. Guyatt GH, Sackett DL, Cook DJ. Users’ guides to the medical literature. III. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993;272:2598-601.
30. Shapiro PE, Pinto FJ. The histologic spectrum of myco-


100. Stem RS, Bigby M. An expanded profile of cutaneous reactions to nonsteroidal anti-inflammatory drugs: reports to a specialty-based system for spontaneous reporting of adverse reactions to drugs. JAMA 1984;252:1433-7.


therapeutics or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994;271:59-63.

The Evidence-Based Medicine Working Group recommends that clinical trials worthy of reading are randomized controlled trials that report clinically important outcomes and that account for all patients entered into the trial. They recommend that articles be read with the following three questions in mind: Are the results of the study valid? What were the results? Will the results help me in caring for my patients? The utility of the method was verified.by comparing medical students who were taught the method with a control group of students. (Bennett KJ, Sackett DL, Haynes RB, et al. A controlled trial of teaching critical appraisal of the clinical literature to medical students. JAMA 1987;257:2451-4.)


5. Chalmers TC, Smith H Jr, Blackburn B, et al. Method for assessing the quality of a randomized control trial. Controlled Clinical Trials 1981;2:31-49. This detailed format and checklist for assessing clinical trials results in the generation of a quantitative score, "index of randomized controlled trial quality," that is a weighted sum of the criteria that are divided into three broad categories: (1) the trial's design and protocol (60%), (2) statistical analysis (30%), and (3) presentation of results (10%).