Acta Derm Venereol (Stockh) 1989; Suppl 144: 15–19

Atopic Dermatitis: Elements in Clinical Study Design and Analysis

BEATRICE B. ABRAMS

Dermatology Medical Research, Hoechst-Roussel Pharmaceuticals, Inc., USA

Considerations in design and analysis of clinical trials in atopic dermatitis are discussed. Since studies analyzed statistically provide an impression of the "probable" effect on the "average" patient, the value of the conclusions depends on the limits imposed by the investigator(s) on factors such as sample size, heterogeneity of the patient population, relevance of the parameters measured and biases introduced in data collection and management. Suggestions are provided for inclusion/exclusion criteria, variables to be measured, sign/symptom scoring systems and data presentations in studies involving patients with atopic dermatitis.

Beatrice B. Abrams, PhD, Dermatology Medical Research, Hoechst-Roussel Pharmaceuticals, Inc., Somerville, NJ 08876, USA.

Clinical trials can be conducted for many reasons; the success of any trial initially depends on a clear definition of the objective and ultimately on the use of proper techniques to achieve the stated goals. Studies can be conducted to gain an early clinical impression of the efficacy of a treatment. In such cases, small populations and uncontrolled designs can be adequate. While the results from such trials may provide the rationale to pursue, discontinue, or modify a specific compound, small sample sizes and nonstatistical design usually preclude generalizations of the results to larger populations. The results will illustrate how the drug will affect the disease in specific patients under certain conditions. Conclusions will depend heavily on the attitude and experience of the investigator and the specific patients placed in the study.

In the development of drugs for market, or comparisons of different treatments, study designs must be developed which will provide data that can be extrapolated to large populations. The conduct of such trials requires relatively large sample sizes plus sophisticated statistical design and analysis. Conclusions based on such results/analyses will provide an impression of the "probable" effects of a new drug on the "average" patient. The results of these trials, however, are meaningless unless the limits of generalizations are set forth in detail. This paper addresses some of the elements of design and analysis which are important in defining these limits, especially as they pertain to the study of drugs for atopic dermatitis.

CLINICAL TRIAL DESIGN

Clinical trials whose results are to be generalized should be randomized, controlled and blinded and the objective should be clearly stated and not overly complex. Parallel group designs usually are required by the Food and Drug Administration in the USA for studies substantiating the safety and efficacy of a treatment; bilateral paired comparison designs are useful in early studies and in some comparative trials. Within this broad framework many types of designs can be employed (see, for example (1, 2)). During the initial considerations of study design, an estimation of the sample size required to achieve the objective should be made. Use of inadequate numbers of patients can lead to uninterpretable data and/or erroneous conclusions. Statistical methods exist which can provide estimates by accounting for the expected variability in the data, the possible magnitude of the differences between the treatments being compared, and the possible numbers of patients who might drop out of the study or be invalidated (3).

The use of multiple study centers helps to prevent a bias introduced by any single investigator and often is necessary in order to provide an adequate number of patients. However, a balance must be achieved between the use of too many centers, which may increase variability, and the use of too few centers, which may decrease generalizability. The use of an odd number of centers provides for a "tie-breaker" if data trends from different centers are contradictory, and scattering the centers over diverse geographical regions and environments helps to offset effects of climate, environment or culture.

Once the study objectives have been established, the variables which might affect study outcome should be defined and decisions made as to what limits, if any, should be imposed to increase the accu-

Table I. *Considerations in clinical trial design: minimizing variability/increasing definition*

Consideration	Control measure
Characterization of condition	Inclusion/Exclusion criteria
Control of concomitant disease/treatments	Inclusion/Exclusion; Restrictions
Evaluation of patient "demographics"	Inclusion/Exclusion; Data collection
Control of dosage form, regimen, duration	Protocol design
Evaluation/control of environmental factors	Protocol design; Data collection
Control of protocol adherence	Instruction; Information cards; Reminder diaries; Drug accountability
Definition of goal	Protocol design

racy and interpretability of the data. Tables I and II list some of considerations in the design of a clinical trial which help to define the limits of the indication/population being studied. The literature abounds with arguments on the pros and cons of carefully defined populations since "overselection" could introduce a bias into the study; however, without some sort of control, data generated will be impossible to pool and treat statistically. If different subsets of the population are expected to react differently to a given treatment—e.g., pediatrics vs adults, stable vs. flaring patients—subject stratification procedures can be employed during enrollment and randomization. However, this tactic is useful only if sufficient patients are available in the different strata to permit meaningful analyses. Unique subsets of the population always can be studied at a later time in separate studies.

Table II. *Critical inclusion/exclusion criteria in atopic dermatitis*

Clear diagnosis of atopic dermatitis, present at least one year
Current flare stable or slowly worsening for more than one week
Lesions suitable for evaluating response to test agents: severity of disease at target site must be such that the total of the numerical ratings for erythema, induration, pruritus is at least 5 out of a possible 9, with *all* parameters being present (rating scale 0=none to 3=severe)
No abnormal clinical, physical or laboratory findings
No hypersensitivity to test medications
No history of alcohol or drug abuse
No concomitant medications during study such as retinoids, antibiotics, large doses of antihistamines, tranquilizers, tricyclic antidepressants
Suitable wash-outs for retinoids, experimental drugs, corticosteroids

Table III. *Definition of efficacy in atopic dermatitis (sign/symptoms)*

Severity on 4 point scale (0=none to 3=severe; half values permitted)
Target area for close observation
Key signs/symptom (erythema, induration, pruritus) *all* must be present in target area; others (lichenification, vesiculation, crusting, oozing, scaling, etc) evaluated as applicable
Global evaluation for total picture (e.g., 0=100% resolution; 1=75% to 99% clear of signs and symptoms; 2=50%–74% clear; 3=25%–49% clear; 4=<25% clear; 5=exacerbation)

Decisions must be made even in the early stages of protocol design concerning data collection and expression. If such considerations are left until the study is completed, important information may be lost inadvertently. The primary variables to be studied must be defined and the manner in which they will be assessed must be determined (Table III). For topical drugs in atopic dermatitis, we generally specify target lesions on which close observations of signs and symptoms will be made; a global evaluation is used to account for changes in the overall condition of the patient's disease. We define the most critical and common signs and symptoms of disease as key signs and symptoms; these must be present in the target area of all patients. Tables IVA and IVB illustrate the prevalence—and severity—of various signs/symptoms during two separate studies of patients with atopic dermatitis (4). In both studies, the most common, and key signs/symptoms, were erythema, pruritus and induration. All others—lichenification, vesiculation, crusting, oozing and scaling—occurred with various frequencies among the population. Since study of these signs/symptoms could be informative, we evaluate them in the target lesion, collect the data, and analyze it as a supportive measure. The lack of adequate numbers of subjects with these signs/symptoms often precludes meaningful statistical analysis.

DATA ANALYSIS/PRESENTATION

Data collected from large studies require careful scrutiny to ensure that data base generated is a correct representation of what transpired during the clinical

Table IV. *Signs and symptom scores frequency distribution at baseline (all subjects in efficacy analyses)*
Scores: 1 = none, 2 = mild, 3 = moderate, 4 = severe, 5 = very severe. Triam = triamcinolone acetonide

Sign or symptom	Treatment	Score 1	2	3	4	5	*N*
Table IVa							
Pruritus	HOE 777	2	0	35	24	5	74
	Vehicle	1	13	27	19	6	68
Erythema	HOE 777	1	17	49	7	0	74
	Vehicle	0	25	32	9	0	68
Scaling	HOE 777	4	27	36	7	0	74
	Vehicle	2	20	39	5	0	68
Thickening	HOE 777	11	17	36	10	0	74
	Vehicle	7	21	31	7	0	68
Lichenification	HOE 777	25	0	29	12	0	74
	Vehicle	21	5	33	7	0	68
Vesiculation	HOE 777	61	10	2	1	0	74
	Vehicle	50	9	7	0	0	68
Oozing	HOE 777	64	15	3	2	0	74
	Vehicle	46	11	6	3	0	68
Crusting	HOE 777	62	14	0	2	0	74
	Vehicle	44	13	7	2	0	68
Table IVb							
Pruritus	HOE 777	1	7	20	25	3	56
	Triam	0	5	20	30	3	58
Erythema	HOE 777	1	6	31	13	5	56
	Triam	0	8	25	24	1	58
Scaling	HOE 777	0	6	27	19	4	56
	Triam	0	3	33	19	3	58
Thickening	HOE 777	1	11	23	17	4	56
	Triam	1	9	20	25	3	58
Lichenification	HOE 777	5	9	24	13	5	56
	Triam	1	8	26	21	2	58
Vesiculation	HOE 777	23	12	15	5	1	56
	Triam	26	14	16	1	1	58
Oozing	HOE 777	30	13	8	4	1	56
	Triam	31	19	6	1	1	58
Crusting	HOE 777	25	15	11	4	1	56
	Triam	29	15	12	1	1	58

trial. The data base also must undergo a certain amount of "cleaning" to define the population to be studied for efficacy. Such cleaning removes patients whose data, for various reasons, might not be a valid representation of the effects of treatment. For example, patients with significant dosing violations should be excluded; data from patient visits exceeding a specified time range should be excluded for the specific visit; data from patients who do not meet protocol entry requirements, used proscribed concurrent medications or were involved in other protocol violations also should be removed. While such "clean-up" rids the data base of interfering variables, it also can introduce a bias. Therefore, results for the "efficacy" population should be compared with those for the entire population and the reasons for any discrepan-

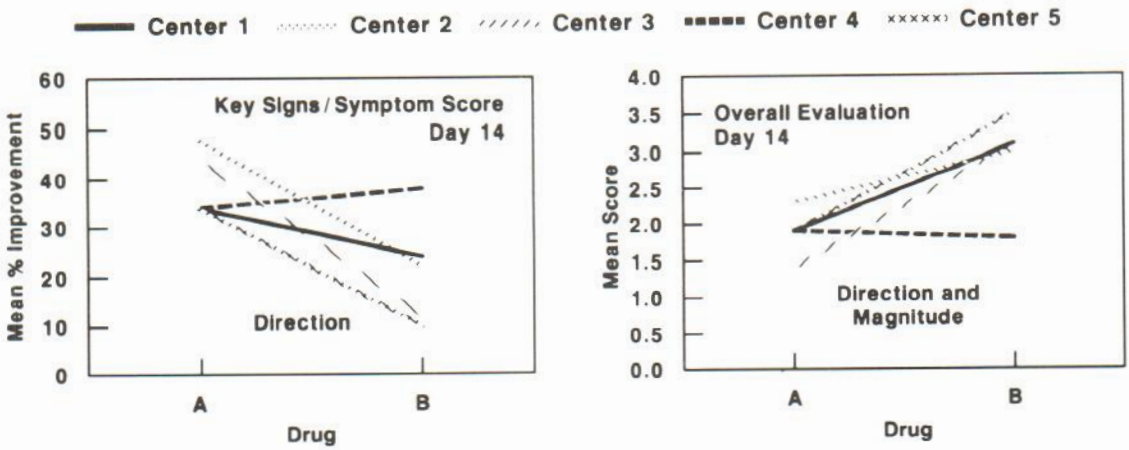

Fig. 1. Results from five centers participating in a multicenter trial comparing two topical corticosteroids in atopic dermatitis. Different lines represent results from different centers. Direction of slopes of lines reflects which treatment (Drug A or B) was better; angles of slopes reflect the magnitude of differences between treatments. (A) Mean percent improvement in key total sign/symptom severity scores after 14 days treatment. Key sign/symptoms (erythema, induration, pruritus) each were graded 1 = none, 2 = slight, 3 = moderate, 4 = severe. (B) Overall evaluation scores: 1 = > 76% clinical improvement; 2 = 51%–75% improvement; 3 = 26%–50% improvement; 4 = ⩽25% improvement; 5 = exacerbation.

cies in outcomes of the two analyses should be investigated. Analyses from the final visit of the efficacy population and the last visit—whenever it occurred —both in the efficacy population and whole population also should be compared as an aid to detect problems which might be causing early termination of patients. Analysis of drug safety data—determination of reasons for dropouts, evaluation of adverse experience reports and analysis of laboratory data—must be performed for the entire patient population, i.e., all patients who received any treatment.

Once the data base is established, a number of analyses can be performed. Obviously, analyses that provide answers to the questions stated in the study objectives should be given primary consideration. The mean change from baseline of sign/symptom severity scores provides a good comparative efficacy measure since comparison of mean scores alone is misleading if baseline scores in the comparative treatment groups are not equal. The severity scores for the key signs/symptoms (erythema, induration, pruritus) can be totaled and the mean change from baseline and mean percent improvement can be calculated to provide an overview of drug efficacy on the major signs and symptoms of disease. Comparison of results of improvement in individual sign or symptom scores and total key sign/symptom scores can be informative for defining the differential activities of an agent. For example, some drugs might be extremely antipruritic, but poor antiinflammatory agents. In such cases, the score for pruritus would reflect dramatic changes; that for the key sign/symptom score would not be as notably affected.

Mean global scores at each visit provide an indication in the change of the overall condition of the patient for all lesions treated. Comparison of these results with those at the target lesion provides an indication of whether the results from target lesion were in fact a valid model for the drug effects on the disease.

While mean values are useful for compressing many results into a single value, they can be misleading since without "qualifiers": they cannot reflect the variability in the measurements made and the range of the values included. Results from three placebo (vehicle) controlled topical corticosteroid clinical studies in atopic dermatitis are presented in Table V (5, 6). In all cases, pooled mean results from all centers participating appear better for the active drug than for the vehicle. However, when one examines the ranges of mean values from the different centers, many of the results for the vehicle and active drug show a great deal of overlap. Because of the deceptive nature of a mean value, comparisons of means should be accompanied by indications of their predicative value such as statistical power statements, standard error values or confidence intervals. Tables exhibiting the distributions of scores also are invaluable for as-

Table V. *Responses to therapy in atopic dermatitis corticosteroid studies. Overlap and range of scores for active drugs and vehicle controls*

Drug	*N*/centers	Day 7	Day 14	Day 21
Poole mean percent improvement in key sign/symptom score 4.6 (range of score at individual centers)				
Active	74 5	45 (31–60)	60 (49–72)	–
Placebo	66 5	25 (6–37)	33 (13–61)	–
Active	89 4	–	–	76 (71–82)
Placebo	90 4	–	–	44 (24–54)
Active	51 3	–	–	84 (72–92)
Placebo	52 3	–	–	23 (7–39)
Global evaluation (1 = none; 2 = mild; 3 = moderate; 4 = severe; 5 = very severe)				
Active	74 5	2.5 (1.0–3.0)	1.9 (1.4–2.3)	–
Placebo	66 5	3.2 (2.8–3.8)	2.9 (1.9–3.6)	–

sessing the full spectrum of results. Such distributions can be complete or grouped. In the former case, using global evaluation scores as an example, one would present numbers of patients with each global improvement score. In the latter case, one could group the scores into clinically meaningful groupings: patients clear vs all other ratings; patients with scores reflecting >50% improvement vs. patients with all other ratings.

Data also must be analyzed to determine whether there were any heterogeneity between the treatment group populations at baseline. The populations' distribution of race, age, sex, and disease severity, duration and state at baseline should be compared as appropriate; the impact of any differences on the study outcome must be determined. Also of importance is an analysis to detect data interactions. In multicenter trials, for example, analysis of variance can be used to determine if results varied significantly at the different study centers involved. Fig. 1 illustrates how results from different study centers can present different pictures of drug efficacy. In this multicenter corticosteroid study in atopic dermatitis, the mean improvement in key sign/symptom scores was better at Day 14 for drug A than for Drug B at four centers; however, at one center, results for drug B were better than those for drug A. Overall evaluation scores showed similar discrepancies between centers. Bias introduced by some investigators (inadvertently, of course) can affect the validity of conclusions from pooled data. The cause or significance of such data interactions must be assessed.

CONCLUSIONS

The basic questions in drug-related clinical research —does the drug work and if so, how does it compare to known standards—often are unanswerable after review of more than one study in the literature. The problem is not necessarily a reflection of poor clinical practice by the investigators. Generally, after careful evaluation, it can be related to such factors as a poor definition of the population studied, variability of the data collected, and/or inappropriate sample sizes studied. Clinicians should be aware of the numerous sources of error and variability in clinical trials when making conclusions from study results, and should insist that clinical trials presented in journals contain adequate information to define the limits of the generalizations/conclusions being made.

ACKNOWLEDGEMENT

I would like to thank Dr Jon Hanifin for his help and encouragement in preparing this manuscript.

REFERENCES

1. Peace KE, ed. Biopharmaceutical statistics for drug development. New York: Marcel Dekker, Inc., 1988.
2. Spriet A, Simon P. (Translated by Edelstein R, Weintraub M). Methodology of clinical drug trials. Basel: Karger, 1985.
3. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 1981; 2: 93–113.
4. Data on file, Hoechst-Roussel Pharmaceuticals, Inc., Somerville, NJ.
5. Data on file, Hoechst-Roussel Pharmaceuticals, Inc., Somerville, NJ.
6. From Medical Officer's Review of NDA 19-543, SCH-32088 Ointment, 0.1% and NDA 19-625, mometasone furoate cream, 0.1%. United States Department of Health and Human Services, Food and Drug Administration Freedom of Information Files F87-19354 and F87-19355.