

SECTION I

CONTINUED

DR. GERALD GUEST, FDA: I'd first like to say on behalf of the Bureau of Veterinary Medicine, we're delighted to be here and to be a part of this Symposium, and to co-sponsor it along with the other groups that are involved. It's an important issue for us obviously and when we asked in the Bureau if anyone was interested in coming, we received an overwhelming response as you can see -- we have a lot of BVM people here because they expect to hear good things and answers to what is probably the only problem we have left; anyway, I think we're down to just this one problem. So Tom, if you'll solve this one tomorrow afternoon, it's going to be nice working around the Bureau. It's a pleasure for me to chair this session; I looked at the three speakers that are on this section and I think without a doubt that I probably have the very best section that will be presented during the program.

We have some interesting things to talk about so I'll stop now and introduce our first speaker. Dr. Jean Powers, as you all know, comes from The Ohio State University. She received her Masters in Genetics there and then went on to get a Ph.D. in Biostatistics -- and that's what we know her best for. She's now jointly appointed in the Department of Statistics in the College of Mathematics and Physical Sciences as well as the Department of Veterinary Physiology and Pharmacology in the College of Veterinary Medicine. Jean is very well qualified to talk on the subject she has today, and I in particular would like to welcome her. Jean's topic is, "What is Statistical Significance in Dose Determination?"

DR. J. POWERS: I think that every statistician in the audience today will agree with me when I say that I am in a very enviable position. How many times does a statistician come before a group like this one and have a predecessor who has presented a lot more formulas and a lot more mathematics than I intend to present? Thanks, Dwight.

What is Statistical Significance in Dose Determination

J. D. Powers, Ph.D. and T. E. Powers, D.V.M., Ph.D.

When a biostatistician, medical scientist or researcher hears the words "statistical significance," the thoughts of rejecting a hypothesis at a given level enters their mind. In addition, the statistician wonders about the "power of the test" or Type II error. Let us define the terms level of significance (Type I error) and Type II error so that we can refer back to this in the light of dose determination. Specifically we have:

1. Type I error = level of significance = probability of rejecting the hypothesis when it is really true.
2. Type II error = probability of not rejecting the hypothesis when it is really false.

Type I error, the level of significance of the test, is very familiar to every researcher. On the other hand, there is a tendency to neglect Type II error. It should be remembered that for a given sample size once the level of significance is fixed, the investigator is "stuck" with Type I error which depends upon the alternative hypothesis. Stated in a different way, for a given sample size both Type I and Type II errors cannot be made arbitrarily small.

The researcher should also be reminded of the problem of statistical versus biological significance. For large sample sizes, a statistical significance could be detected which has no biological meaning, e.g., a group mean of 120 mm Hg systolic pressure could be shown to be significantly smaller than 123 mm Hg if "n" is large enough--this is not important biologically.

Now we shall turn our attention to how a dose of a drug is determined. For both NADA's and NDA's the procedure is roughly divided into two parts: preclinical trials and clinical trials. Initial dosage determinations are carried out in the preclinical trial investigations and then "verified" or adjusted during or immediately following the clinical trials.

The entire drug approval process, at least from my point of view, is entered around the statement:

"A drug must be shown to be safe and effective in adequate and well controlled studies."

Again, as with the definition of Type I and Type II error, let us look at this statement more closely so that as the discussion continues we have a common point of reference.

1. Safe - All one needs is to open a PDR or VPB and immediately it is clear for literally every product listed there are contra-indications and possible side effects. Is it "safe" to give chloramphenicol to humans when it has been shown to cause aplastic anemia? It is given, however.
2. Effective - Does this refer to pharmacologic or therapeutic efficacy or both? Further, how does one measure therapeutic efficacy--certainly dichotomous ratings such as live/die or improved/not improved leave much to be desired from a quantitative point of view. If therapeutic efficacy is measured on a more detailed scale, then the question of inter-rater reliability is raised. One cannot discuss efficacy without being reminded of the story about the patient being treated with a new chemotherapeutic agent. Pharmacologically, it was very efficacious--the tumor was beginning to atrophy, however, therapeutically the drug was far from efficacious--the patient died from secondary invaders. Or conversely all of you have seen cases of the poor sick animal getting well with only a placebo and TLC.
3. Adequate - This is possibly the most ambiguous term of them all. Is it "adequate" to show a drug is safe and effective in 10, 50 or 100 patients? Is it adequate to show it is effective and safe "most of the time?" Or does adequate refer to sample size or number of trial sites?
4. Well controlled - What is well controlled and specifically how many control groups are needed. What are the options?
 - a. Historical control. Clearly no untreated control group is needed to predict the outcome of the patient bitten by a rabid animal. However, consider the Boston surgeon who showed favorable results when he shunted patients with cirrhosis while the control group had a markedly worse survival. These were historical controls. It was claimed he operated on patients with a good prognosis and not on those with a poor prognosis. Not so, he said, he never selected his patients, he only selected the time of surgery. In other words, he only operated on those who lived long enough to get into good enough shape for surgery, and he compared these with a group which included some who did not live long enough to get into good enough shape for surgery.
 - b. Untreated control. This type of control is important to determine if the animal will get well on its own. However, is it ever ethical in either human or veterinary medicine to allow patients to go untreated, especially in life-threatening situations.

- c. Healthy control. Is the physiology of the healthy subject comparable to the "sick" patient or are we comparing the proverbial apples and oranges?
- d. Vehicle control. Is this control group needed if the vehicle is water or saline. Probably not. What if the vehicle is propylene glycol?
- e. Current drug of choice control. Is the "new drug" equally effective as one which is presently available for use?

With these definitions and the statement as to what must be shown before us, let us turn our attention to the question at hand, "What is statistical significance in dose determination?"

Preclinical Investigations

It is during this phase of drug development that the "proper" dose is established. Most will agree, if I may use a statistical term, there is not one unique dose. Rather, there is an entire spectrum of doses ranging from a minimal dose which is effective to a maximal dose which is nontoxic. From a statistical viewpoint, a minimal effective dose must be shown to be the lowest dose which is effective. Likewise, a maximal nontoxic dose must be shown to have no toxicities as compared to the next higher dose. This concept causes concern to a statistician because neither of these doses can be determined uniquely. This can be best described by an example: Suppose a drug is tested at 10, 20, 40, 80, and 160 mg/kg. Suppose from the results of the investigation the researcher would like to claim that 20 and 80 mg/kg are minimal effective and maximal nontoxic doses, respectively. Statistically this would be a valid statement if the investigator makes the qualification that among the five doses tested, those two have the properties described. This is an example in which Type II error should be addressed.

On the surface, it would appear the criticisms of lack of uniqueness of a specific dose could be minimized by including more doses, i.e. narrow the interval between adjacent dose levels. This is only one aspect of the problem. Let us confine our discussion to antibiotics. For any drug involved in an NADA or NDA, more than one species of pathogen will show susceptibility. Hence, one is potentially dealing with a series of minimum inhibitory concentrations (MIC). Indeed not only are the MIC's different but also the pathogens could be located in many different tissue sites all with varying coefficients of permeability.

As stated earlier, it is during this preclinical phase that investigations are designed and carried out to establish a "proper" dose to be used during the clinical trials. How does one design such an investigation? Let us describe a possible design which would be "ideal" from a statistical point of view. The following factors should be included:

1. Dose levels - at least 5

2. Species of pathogens - maybe 4
3. Sites for each pathogen - at least 4
4. Ages of patients - 3 (neonate, adult, geriatric)
5. Route of administration - 3, e.g., oral, IV, IM or if only an oral preparation is to be marketed, the 3 groups could be tablet, capsule and enteric coated time release product.
6. Dose interval - 3 (BID, TID, QID)

If one assigns only 6 subjects per group, it would require 12,960 animals. Clearly this is an unrealistic undertaking and compromises are necessary. Suppose only one route of administration is tested, even then 4320 subjects are needed. At this point, let us assume that preclinical investigations have been completed and the drug is ready for clinical trials.

Clinical Trials

The statistical considerations involved in clinical trials are totally different from those in the earlier investigations. Any list of points to be addressed would include the following:

1. Types of controls
2. Stratification
3. Compliance
4. Premature exit from study
5. Patient selection and exclusion criteria
6. Variables to evaluate therapeutic efficacy
7. Ethics

Earlier we described the different types of controls which could be used. We are all cognizant of many examples of the "placebo effect," but for the sake of completeness one will be included. The veterinarian is acutely aware of the therapeutic effects, i.e. placebo effect, of "tender loving care." Not long ago, an investigator at Ohio State University fed two groups of rabbits identical high cholesterol diets. One group was managed according to GLP instructions. The other group was managed in the same fashion except at regular daily intervals the animals were removed from their cages and petted. Later this latter group was shown to have statistically fewer aortic atherosclerotic plaques. Interesting enough, this was shown with relatively small sample size.

As a statistician I would be remiss if I did not emphasize the importance of randomization of subjects to treatment groups. Randomization could mean to all subjects in general or it could mean to all subjects within a stratum. The design of the experiment (trial) dictates the level of randomization. One cannot emphasize too strongly the need for statistical "blinding" in the case of clinical evaluators. All too often the evaluators consider "blinding" a form of statistical game playing. The evaluator is a professional clinician and because of this either conscientiously or subconscientiously has ideas about which treatment is the best. This type of bias must be minimized at all costs.

Diagnostic selection criteria need to be employed in clinical trials however if these criteria become too selective or restrictive, the purpose of the trial is defeated, i.e. to apply the results of the trial to clinical medicine. Let me take an example from human medicine to illustrate stratification. Suppose we wish to evaluate the efficacy of an acid secretion, blocking agent on the healing of gastric ulcers. There are three types of ulcers. 1) those in the gastric body, 2) those near the pylorus and 3) those combined with the duodenal ulcer. Since antisecretory agents are not equally effective against all three types, one may wish to stratify on ulcer type. Statistically, the effect of stratification is to decrease or minimize within group variability which allows the detection of "smaller" differences.

Compliance is always a problem to be addressed except in the case of the hospitalization of all subjects on the trial or in the case of the trial in which the "treatment" is administered by the clinician. It is necessary that each investigator do everything possible to ensure compliance.

The patients who are "lost to follow-up" present statistical problems however these problems are tractable. The investigator is charged with the task of educating and managing his subjects so as to keep this problem at a minimum.

Patient selection and exclusion criteria must be held in a delicate balance. If the selection and exclusion criteria are too restrictive the clinical applicability is in jeopardy. On the other hand if there are no selection or exclusion criteria, the variability between and within subjects will be so large as to make it very difficult to detect all differences except those so large as to be obvious.

How do we evaluate therapeutic efficacy? What variables do we use to measure efficacy? It is imperative to have a scale which a) evaluates the "state of health" of the patient and b) is consistent across investigators.

The question of ethics has been left as the last point of discussion because it presents a problem for the investigator which only he can resolve within his own conscience. Is it ever ethical to withhold treatment (placebo)? Finally is it ever ethical to require necropsy reports for all groups in a laboratory study because it is an animal. Those

investigators and evaluators in studies on humans obviously do not have that requirement. The privilege of carrying out studies involving the use of animals will only be continued if the privilege is not abused. Under the present "moral" conditions there are groups of people looking for reasons to revoke this privilege. From my personal observations a few of their concerns are justified. Let us suppose you are attempting to get an antibiotic approved for use in horses. For such a NADA in either the pre clinical or clinical trials there could be two controls, current "drug of choice" control and the healthy control. I believe it is a flagrant misuse and abuse of animal life to require a necropsy for either of the groups. In the former case an approved drug is being used. This drug has already "passed" the necessary testing criteria and a necropsy report is a re-evaluation of an accepted drug. In the latter case, to require necropsy of healthy controls is a mockery of both the veterinary clinician who judged the animal healthy before assigning it to a group and the veterinary pathologist who certainly knows what healthy tissue looks like. We have all heard the somewhat facetious claim that if all regulatory criteria and guidelines are met, more animal lives will be lost than could possibly be lost in the clinical situation. I know animal welfare is very important to the academic veterinarian and earlier we heard Dr. Harvey express the same concerns of the FDA, hence there is no reason why these concerns cannot be met.

There are many other issues over which the investigator has little or no control. However these same issues can influence the outcome of the trial. To name a few of these points:

1. Comparable nutrition for all subjects
2. Living conditions
3. Time from onset of disease to entry into study
4. Exposure to other animals

After the completion of the clinical trials, the final step before submission is the statistical evaluation of the results.

At this point I have no intention of describing statistical techniques commonly in use today. The statisticians among us are fully aware of the tests and the veterinarians and lawyers among us are quite willing to leave it to the statistician. Let me instead just describe one technique which has been used only in a limited number of investigations involving NADA. This technique is called sequential analysis. The statistical literature is filled with examples and indeed entire text books have been written on the subject. Simplistically speaking a trial or test is begun with stopping criteria having been defined earlier. The data are evaluated at predetermined intervals and the test or trial is discontinued when a) a stopping criterium is met or b) a maximum number of animals have been included.

This type of testing procedure has several obvious advantages:

- a. potentially saves time
- b. potentially saves money
- c. Most importantly minimizes the number of animals necessary for testing.

It makes "good sense" to maximize the informational value from a minimum number of subjects. There are also precedents to which we can refer. Within the past year, the public media informed us that only a few years into a twelve year trial the testing was stopped because of overwhelming evidence to that date. The very cautious among us will remind us of some of the catastrophies which may be averted by prolonged testing, the prime example is always thalidomide or more recently the organo-phosphate problem in puppies with respect to the former teratogenic studies have become common place. With respect to the latter - careful assimilation of information at hand may have averted the problem, e.g. we are warned to not "over stress" our horses immediately following such a treatment, that should have alerted us to the potential toxicities when used in immature subjects. To date the "perfect experiment" has not been designed, we can only minimize the risk of a problem by maximizing the utility all other information available to us.