

Statistical View of Clinical Trials: An FDA Perspective

James J. Colaianne, Ph.D.*

This morning has been very interesting. We started out with Dr. Tom Powers' introduction and I was feeling very good about my talk because Tom said we would have a chance to edit all our comments and deny everything we said. Then Bob Griffith spoke and I realized that he was saying about everything I intended to say. And finally Dan Gingerich spoke criticizing clinical trials and now I am not sure I have anything left to say at all. But assuming that we are going to continue to do some kind of clinical trials, what I want to do is give you an outline or overview of the kind of things the FDA, and the CVM statistician look for. Of necessity I am not going to be able to hit all of the critical areas, but I will try to point out those places where we run into our biggest problems.

It was mentioned earlier that one of the major motivations for this symposium has been the frustration experienced by FDA as well as the drug industry, when a drug sponsor submits clinical trials with lots and lots of case reports, only to have FDA turn around and disqualify the vast majority of them. There are usually a number of specific reasons for these disqualifications and Bob Griffith has covered some of them. My point however is that there are also some basic underlying causes. In my

*Chief, Biometric Staff, Scientific Evaluation, Bureau of Veterinary Medicine, Food and Drug Administration, 5600 Fishers Lane, Rockville, Maryland 20857

opinion, the major underlying reasons are inadequate planning, poor experimental design and sometimes lack of monitoring. The statistician can be quite helpful in avoiding problems, particularly in the first two areas, provided he is brought in early enough. He needs to be in at the early planning and development stage as well as the formal protocol development stage. You have probably heard this before, the statistician needs to be in on the entire protocol development. Moreover, it is also a good idea to keep him involved throughout the conduct of the study and certainly throughout the data interpretation phase of the study.

Now let me stop philosophizing and turn to my outline. I am going to try to hit the key points of this outline, some of which you may have heard before. I mentioned many of these points in a paper I gave to a similar group about two years ago at Ohio State. (Colaianne, 1982)

The first and probably the most critical point in a clinical trial is to decide on its purpose. Time and time again, we get clinical trials where the purpose is unclear. It has been our experience, the more precise the purpose, the more precise the definition of the action of the drug, the more focused the definition of the response population and the condition that is going to be tested, the more likely you are to have success. Vague definitions invariably lead to trouble. As an example, we actually do see studies with purposes as vague as .."to show a new drug aids in the prevention and control of respiratory infections.." If you stop and think about it, this is really asking for failure. It is very questionable if one can ever conduct clinical trials to meet such a

broadly stated purpose using reasonable numbers of patients. Maybe that is the basis for Dr. Gingerich's criticism of "controlled" clinical trials, I am not sure. I know we see a lot of studies with these types of purposes at FDA; possibly due to the fact that the market place says a good drug has a wide action and a wide claim. On the other hand, we at FDA tend to limit claims to single species and very restrictive kinds of actions partially due to our concerns that the clinical trials clearly demonstrate efficacy and the need to limit the purposes of the trials in order to obtain this demonstration.

Let me move to the second area of concern the experimental design. Dr. Jean Powers has already mentioned that the design area is probably the most critical area of statistical assistance. At FDA, it is this area where our statisticians spend the majority of their review time and where we make our greatest impact.

In this talk, we do not have sufficient time to cover all the possible types of designs that might be used in a clinical trial but I do want to touch upon probably the most common type of design the randomized complete block design. With this type of design, each investigator would be expected to be treating patients with all of the treatment regimens involved in the study. If we are talking about a simple study that only involves one treatment group and one control, he would be handling patients in both groups. And, ideally he should have equal numbers of patients in each group. The main reason for wanting equal numbers in

each group is that we have had problems arise if an individual investigator gets disproportionate numbers of patients on control or treated. This disproportionality can lead to a bias when the data are pooled from more than one investigator.

As an extreme example consider a case where ten times more patients are placed on the drug treatment than on the control by one of the investigators. If this investigator tends to be rather optimistic in his evaluation of patient improvement, then he will contribute ten times more successes to the drug treatment group than to the control group in the pooled results which could lead to considerable investigator bias in the final evaluation. The safest rule to follow is to try for balance in all areas of the design, within an investigators, within treatment groups, and certainly across the investigators. I realize that this is not always practical. One investigator is going to see a lot more of a particular condition than another. But if one investigator has such a preponderance of the patients, that his results totally overshadows all the other investigators, this could make the results irrelevant with respect to the total population of practitioners. FDA does look for this type of imbalance in clinical trials and may invalidate the findings of the entire study due to extreme imbalance.

One other design I want to touch upon is the crossover design. This design is great for bioequivalency, blood level studies, but it has generally been discredited for clinical trials. The major reason is that in a clinical study it is very difficult to argue that there isn't some type of drug carryover effects once a sick animal has been

treated. One of the most critical assumptions in the crossover design is that the order that the patient receives the drug or control regimen, is not important and does not influence the results. We know that this will usually not be the case with clinical trials where we are treating "sick" animals and where once a patient has been treated with the active drug the patient will usually not relapse to its previous pretreatment condition.

Now I would like to turn to the issue of the selection of a control group in a clinical trial. From the standpoint of a statistician, the controlled clinical trial is an absolute must. As clinicians, and investigators, you may feel control groups are unnecessary and that you can evaluate drug efficacy in an uncontrolled trial. But from my viewpoint as a statistician, I generally doubt the validity of results from such studies. Most of the things that we end up arguing about with the industry have to do with subtle changes, subtle differences in responses, not clearcut kinds of situations where we are counting the numbers of dead animals. With these subtle kinds of measurements we need some kind of baseline or reference point to verify that observed changes and improvements are due to drug response not just spontaneous remission.

There are four typical controls that we frequently see proposed: placebo, untreated, positive and historical. Naturally my favorite is the placebo. There are lots of ethical criticisms associated with the use of a placebo, but it is the most effective and clearest way to test a drug's effect. Moreover, arguments have been made time and again in human medicine that the use of placebo is not only the best type of

control group to use, but provided you are not dealing with a life threatening situation, or a situation where you are allowing excessive pain and suffering (at least within the short time frame of the clinical trial) that it is really more ethical to use a placebo than no control or a historical control, because you will then have a better opportunity of detecting new, important drug entities. I can not deal with the ethical issues in this short talk but let me simply state that from my viewpoint, the placebo control provides the cleanest, clearest kind of comparison we can get in clinical trials.

Now let me turn to untreated controls. These still crop up once in a while in our clinical trials and as you might expect, they are usually associated with trials in food animals. In these situations they are essentially the same as a placebo. We run into many of the same problems, except that they do not have the niceies associated with a placebo where all of the concomitant activity or therapy is represented in the control, with the exception of the absence of the drug. However, even in food animal trials, I would prefer to see a placebo control if I can get one.

Probably the most frequently proposed control for a clinical is the positive control. FDA obviously accepts these and they can be quite useful. However, they do present two major problems. If you anticipate that the new drug entity that you are testing is going to be better than the standard positive control, the chances are that it will take a larger number of animals to show this in the clinical trial than it would take to show an improvement over a placebo control.

The other problem that we typically run into at FDA, is many times we do not know whether the new drug entity is going to be better than standard, it may be only as good as the standard. In this case, use of a positive control really gives you no information at all. If we observe no differences between the patients treated with the new drug and those treated with the positive control we can not really conclude that the two drugs are equivalent since we are often uncertain as to whether this study was really big enough to pick up meaningful medical differences if they existed. In general, if we do some kind of statistical calculation to compare the efficiency of the various types of controls will find that it takes two, or three times as many patients to reach the same kind of statistical confidence using a positive control as you get when you use a placebo or untreated group.

Let me move on now to the historical control only for purposes of completeness. I heard some discussion earlier today as to why FDA does not allow the use of historical controls. The classic answer is that you do not know all the conditions under which the historical results have been measured. In most of the trials FDA faces, we can not be sure that the present is consistent with the past. We are not, at all sure that the historical response associated with a disease is really relevant with respect to the present population being treated with the drug under test.

Let me put it in a little different way. In general we do not accept historical controls but they could be accepted in very rare situations where you have a very dramatic disease condition, a well defined end points with very predictable results if you do not treat. The classic

example is rabies where death is historically anticipated if animals are not inoculated, however, even here I recently found out that death is not certain and results can be equivocal if no control is present. If pressed for a rule, I would state that historical controls are not accepted by FDA.

That takes us to randomization. We at FDA always look very closely for randomization in clinical studies and we continually have problems in this area, even in the food animal clinical trials. Randomization is essential to minimize bias, and to prevent manipulation of the order in which patients are assigned to groups. Manipulation does occur, despite good intentions, and it can lead to very complex bias. As a result complex methods have been developed to accomplish randomization. Dr. Jean Powers has already mentioned that complete randomization is very rarely used. It is not ideal, since you usually have to worry about some degree of balance in the study and you want to randomize within sexes, or within breeds or any of a host of other classifications that might influence the results of the study. Randomization is almost an art unto itself and lots of papers have been written on it, but the basic idea is simple, you do not want unintentional biases to influence the key comparison between the treatment and control groups.

Let me emphasize something I have said in a couple of different places in the past. Good randomization is not haphazard. We are not talking about setting up some kind of random assignment scheme by rolling a die or , flipping a coin. Those methods can be used, but there are far more sophisticated methods that are not that difficult to apply, such as random number generator tables. Instructions are usually provided with

these tables in the back of most standard statistical texts and they are reasonably simple to follow. The critical concern is that use of haphazard randomization (i.e. pseudo-randomization) can actually lead to more systematic bias than the use of no randomization at all.

In a similar vein, I want to emphasize that randomization schemes also need to be unique. Unique in the sense that you do not want to be reusing the same old randomization scheme from one clinical trial to another or even one investigator to another. The justification for this is that despite the efforts to make the randomization scheme truly random, you may have unintentionally introduced some bias or systematic effect. If you keep using the same scheme over and over again, you magnify this bias.

Lastly, let me mention in this area, one specific assignment scheme that keeps cropping up in veterinary medicine. This is the alternating assignment scheme. Alternating assignments is not a randomization scheme at all, but instead is completely systematic. Its intention is to maintain balance in the study. Typically with this scheme the first patient into the study is put on, say, drug the next one on control, the next on drug, the next on control, and so on in an effort to keep the groups balanced with respect to numbers at all times. This is a very admirable goal but this kind of scheme has a very high chance for introducing bias. Again, it may be quite unintentional but an investigator could very easily manipulate which patient gets into which group simply by determining where he is going to put him in the sequence of patients entering the study. Consequently we do not recommend the use of alternating assignment, despite its simplicity and convenience.

Let me move quickly to blinding. I do not have much to say on this subject. Most of you here know more than I do about double and single blinding and the sorts of problems associated with these. Typically, blinding is done at the patient assignment phase of the study, which is excellent. My only pitch here is that, as a statistician, I would also like to see blinding at the patient evaluation phase. We do not see this latter situation very often. However, by blinding at the patient evaluation phase, we might be able to avoid the situation where an investigator does not know whether group A or group B is the test drug group, but he does know whether a patient being evaluated is in group A or B. In this situation he may have a tendency whether he means to or not, to magnify differences by comparing observed results between groups as he evaluates. A blinded evaluator would not know whether the patient is from A or B so he would not be inclined to make this kind of cross comparison.

I have just a few more comments to make with respect to response variables. I believe a number of people have already pointed out the need to clearly identify the response variables in a clinical study. The ideal is to have a single objective response variable as the primary indicator of drug efficacy. In practice we rarely find this. If there is going to be more than one response variable looked at in a study, there should be some kind of hierarchy. Based on the medical situation involved, the investigator should decide which of these is the most critical response variable and which is second. Generally there should be no more than two or three critical variables identified in the study,

despite the fact that it may be possible to measure 10 or 20 variables analytically.

Let me make two quick last points. With respect to data analysis, I think there is a misconception that most of the statisticians in industry and FDA argue over this aspect of the study. Actually we have very few arguments over the data analysis. If the statisticians are allowed to get involved at the early planning phases, both on your side (industry and academia) and on my side (FDA), then the statisticians usually can agree on the methodology. These agreed upon methods can be written into the protocol and any necessary minor adjustments can be made, after the fact, in the actual analysis of the results.

However, I do have a concern about sequential data analysis. The procedure of "taking a look and deciding if you want to keep going" in a clinical trial is really very dangerous. If you are dealing with the comparison of a treatment and a control and they are constantly reversing their position, one to the other, as you add patients, where you stop may determine whether or not you see a difference and the direction of the difference. When you take multiple looks in a clinical trial, you should be adjusting for the fact that you are sequentially evaluating the results. There are statistical analysis methods (sequential analysis) that account for the fact that you are going in and looking at the data, and then making a decision. These types of statistical procedures need to be applied in order to avoid drawing incorrect conclusions.

Lastly, with respect to interpretation of results, generally at FDA we require that you show statistically significant differences between the control and the treated group in clinical trial. But the FDA statistician do not work in a vacuum but in concert with our veterinarians and whether or not a drug is considered to have been demonstrated effective after the statistical analysis has been run will not be just a statistical question. I am sure there are people in this audience who have experienced the situation, where they got a statistically significant ($p < .05$) improvement with a drug but were told the response was not quite dramatic enough to call the drug effective. Statistics is not the final line in interpreting the results of clinical trials.

One final summary statement. In this talk I have tried to point out that a statistician can be very helpful in developing clinical trials and possibly help get away from these high loss rates that many have experienced when the case reports are submitted to FDA. But in order to help, the statistician needs to be involved early in the study and stay involved throughout. You should appreciate that whether or not you involve your statisticians in the planning and conduct of your study, FDA is going to involve theirs in evaluating the acceptability of the design and in evaluating the significance of the results.

Reference:

Colaïanne, J.J. 1982 "Statistical Considerations in the Design of Clinical Trials" pp 37-48 in Topics in Veterinary Medicine ed. by J.D. Powers and T.E. Powers, Ohio State University Press, Columbus, Ohio