Clinical Trials from the Statistician's Side: Reducing Variability

BY ARON SHAPIRO; AND DALE USNER, PHD

We weren't all born with math on the mind. Lucky for us, statisticians help to fill the void that many of us face when it comes to crunching numbers. Statistical reasoning is vital to any clinical trial, as it influences study design, data collection, and data analysis. Put simply, statistics is the art of summarizing data; better yet, summarizing data so that nonstatisticians can make meaningful conclusions. Clinical investigations typically involve collecting large amounts of data, but at the end of the trial, we want the punch-line: Did the new treatment work, and is it clinically meaningful? One key statistical component to any clinical trial is minimizing variability. In this month's column, we look at ways to reduce variability in a clinical trial through a hypothetical study.

REDUCING VARIABILITY

When minimizing variability in retinal imaging data for a clinical trial, there exists a "uniform operations chain" consisting of the machine, operator, and reading center. First, it is important that the machine utilized is up to proper precision; if there are multiple study visits, it is beneficial for each site to consistently use the same machine for each exam. Next, whether the sponsor uses internal staff or contract monitors, it is important to ensure that these individuals are properly trained to use the equipment. It is also helpful for the same personnel to be used throughout the duration of the study, as maintaining consistent staff will help to lessen the subjectivity and variability involved with data collection and analysis. Additionally, standardizing the assessment methodology and grading criteria is of utmost importance. As an effective way to improve precision and reduce variability, it may be useful to create an operations manual that contains specific instructions for imaging and an upfront practicum to ensure that procedures are being followed. Included could be written directions for carrying out every procedure: for example, a standard procedure for imaging and reading the image. Finally, the primary read of the image should be performed at a central reading center. In addition to providing standardized reader training, this will ensure that the process is accurate, and that bias and variability are minimized.

It may be useful to create an operations manual that contains specific instructions for imaging and an upfront practicum to ensure that procedures are being followed.

OUR HYPOTHETICAL STUDY

Setting up and testing hypotheses is an essential part of statistical inference. The question of interest is simplified into 2 competing claims between which we have a choice; the null hypothesis, denoted H_0, against the alternative hypothesis, denoted H_a. Frequently a clinical trial is designed to show superiority of a new treatment (drug/biologic/device) over control (placebo/standard of care). In a superiority design, H_0 assumes that the measure of interest is the same between the 2 treatments, and H_a assumes that the measures are different in favor of the new treatment. Determining the required sample size to test such hypotheses relies on 4 key parameters:

1. The assumed difference in the measure of interest between the treatment groups in H_a; the larger the difference, the smaller the required sample size.
2. The assumed standard deviation (sd) of the mea-

sure; the smaller the sd the smaller the required sample size.

3. The type 1 error rate (α), which is the probability of rejecting H_0 when H_0 is true (ie, the probability that the trial will incorrectly show that the new treatment is better than control); this probability should be small. For pivotal trials used for regulatory approval of a product, α is generally set to 2-sided level 0.05, which means that the test will incorrectly reject H_0 when H_0 is true in 5% of trials. Half of these times, 2.5%, the test will incorrectly reject H_0 when H_0 is true and show that the new treatment is superior control; the remaining 2.5% will show that the new treatment is inferior to control; therefore, in this type of design a 2-sided α = .05 is the same as a 1-sided α = .025. The smaller the α the larger the required sample size.
4. Power, which is the probability of rejecting H_0 when H_a is true (ie, the probability that the trial will correctly show that the new treatment is better than control); this probability should be fairly high, generally 80% or 90% for pivotal trials to maximize the probability that a new treatment that is truly efficacious is shown to be efficacious in the trial. The larger the power the larger the required sample size.

The probability that a trial rejects the null hypothesis assuming the alternative hypothesis is true (power) decreases as the signal-to-noise ratio (assumed difference divided by the sd) decreases. For example, consider a study with the following hypothesis:

H_0: The mean difference between Test and Control in the reduction of central subfield thickness by SD-OCT from baseline to Month 6 = 0 µm

H_a: The mean difference between Test and Control in the reduction of central subfield thickness by SD-OCT from baseline to Month 6 ≠ 0 µm. Where superiority of Test over Control will be concluded if the mean difference (Test – Control) >0 µm.

Consider that previous studies showed a mean change from baseline for Test of 155 µm and a mean change from baseline for Control of 145 µm (ie, an assumed difference of 10 µm), each with a sd of 25 µm, yielding a signal-to-noise ratio of 10/25 = .4. With this signal-to-noise ratio, 200 subjects (100 subjects per treatment group) are required to have 80% power assuming a 2-sided α = .05 test. If there are only enough resources to study 130 patients, 1 way to reduce the number of required subjects and maintain power and α is to reduce the sd.

If the estimate of sd from the previous studies came from local SD-OCT reads at different sites on various machines, where the sites were not trained on a standard procedure for both imaging the eye and reading the image, then future trials should be able to reduce the sd by implementing any of the following: training the sites on a standard procedure for imaging and reading the image; consistently using the same SD-OCT machine at each visit; and/or having the primary read of the image performed at a central reading center. For example, training the sites on a standard procedure would decrease the sd from 25-20 µm. This reduction in sd [and increase in signal-to-noise ratio from 0.4-0.5 (10 µm/20 µm)] would reduce the required number of subjects to 128 (64 subjects per treatment group). Additionally, implementing the use of the same SD-OCT machine at each visit and a central reading center fur-

TABLE 1. SAMPLE SIZE ASSUMING A 2-SIDED α = 0.05 TEST OF STATED HYPOTHESIS

Signal-to-Noise Ratio (Mean Diff/sd)	Example Ratios		Total Sample Size 80% Power	Total Sample Size 90% Power
2	10/5	4/2	12	14
1.5	10/6.667	4/2.667	18	22
1.25	10/8	4/3.20	24	30
1	10/10	4/4	34	46
0.8	10/12.5	4/5	52	68
0.667	10/15	4/6	74	98
0.5	10/20	4/8	128	172
0.4	10/25	4/10	200	266
0.333	10/30	4/12	286	382
0.25	10/40	4/16	506	676
0.2	10/50	4/20	788	1054

TABLE 2. AS THE SD DECREASES (AND THE SIGNAL TO NOISE RATIO INCREASES), THE SAMPLE SIZE REQUIRED DECREASES.

Signal-to-Noise Ratio (Mean Diff/sd)	Proportion > Midpoint		Total Sample Size 80% Power	Total Sample Size 90% Power
	Active	Control		
2	0.84	0.16	16	20
1.5	0.77	0.23	26	32
1.25	0.73	0.27	36	46
1	0.69	0.31	52	70
0.8	0.66	0.34	76	100
0.667	0.63	0.37	114	152
0.5	0.60	0.40	194	260
0.4	0.58	0.42	306	408
0.333	0.57	0.43	400	532
0.25	0.55	0.45	784	1048
0.2	0.54	0.46	1226	1638

ther reduces the sd to 15 µm. This reduction in sd [and increase in signal-to-noise ratio to 0.667 (10 µm/15 µm)] would reduce the required number of subjects to 74 (37 per arm).

In this hypothetical study, the total sample size required for the study decreases from 200 to 128 to 74 subjects as the signal-to-noise ratio increases from 0.4 to 0.5 to 0.667 (sd decreases from 25 to 20 to 15 µm) through implementing standardized imaging and reading procedures (Table 1). Implementing strategies to minimize the variability in the measure of the primary endpoint substantially reduces the number of subjects required, the cost, and the duration of a clinical trial and should therefore be investigated.

Another analysis method frequently used in clinical trials is to dichotomize a continuous measure into two groups: 1) those subjects whose measure is greater than a specified value, X, and 2) those subjects whose measure is less than or equal to a specified value, X. This new measure is then analyzed to determine if the proportion of subjects whose measure is greater than X is different between the two treatment groups. However, care should be taken with dichotomization of continuous measures, as generally the sample size required to show a difference in the proportions is greater than the sample size required to show a difference in means on the original continuous scale.

Continuing with the example above, assume that the change from baseline in central subfield thickness scores follows a normal distribution with the same sd for each treatment group, and define a dichotomous Yes/No variable of the form:

1 if the change from baseline is > X
0 if the change from baseline is ≤ X

Defining X as the value at the midpoint between the means of the treatment arms will yield the highest probability for detecting a difference between treatments of any Yes/No variable of the form. In our example, the midpoint between the means of the treatment arms is (155+145)/2 = 150. Therefore, testing the difference in the proportion of subjects with a change from baseline greater than 150 µm, through the hypothesis below, yields the highest probability of detecting a difference between the treatments using a dichotomous variable.

H_0: The difference in the proportion of subjects with a change from baseline to Month 6 >150 µm in central subfield thickness between Test and Control = 0
H_a: The difference in the proportion of subjects with a change from baseline to Month 6 >150 µm in central subfield thickness between Test and Control ≠ 0.
Where superiority of Test over Control will be concluded if the difference in proportions (Test – Control) >0.

For example, an expected proportion of:

Test subjects to have a change from baseline >150 µm is
- 58% (assuming a sd of 25 µm)
- 60% (assuming a sd of 20 µm)
- 63% (assuming a sd of 15 µm)

Control subjects to have a change from baseline >150 µm is
- 42% (assuming a sd of 25 µm)
- 40% (assuming a sd of 20 µm)
- 37% (assuming a sd of 15 µm)

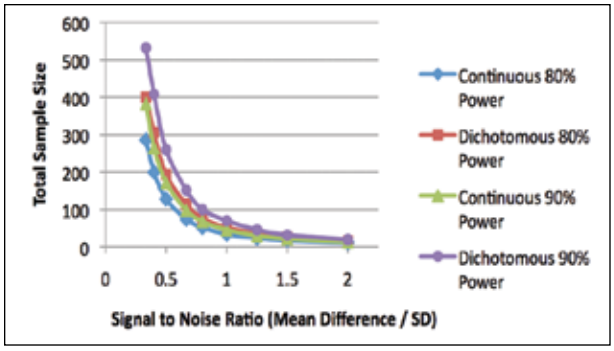

Figure 1. Depicting the sample size requirements by signal-to-noise ratio as shown in Tables 1 and 2.

Note that as the sd of the measures on the continuous scale decreases, the expected proportion of change from baseline scores greater than 150 µm increases in the Test treatment subjects, the expected proportion decreases in the Control treatment subjects, and the difference in the expected proportion increases. Therefore, as shown with tests on the continuous scale, as the sd decreases (and the signal-to-noise ratio increases), the sample size required decreases.

In this hypothetical study, using the dichotomous endpoint, the total sample size required for the study decreases from 306 to 194 to 114 subjects as the signal-to-noise ratio increases from 0.4 to 0.5 to 0.667 (sd of continuous measure decreases from 25 to 20 to 15 µm; Table 2). Therefore, reducing the variability of the measure also reduces the sample size required when dichotomizing the measure. However, converting from a continuous to a dichotomous variable requires from 33-55% more subjects to test the differences between treatment groups (Figure 1).

CONCLUSION

Conducting a clinical trial and interpreting the results are complex, involved processes. It is important to keep in mind the measures of clinical significance, and the difference between *statistically significant* and *clinically meaningful*. Statistical significance does not necessarily translate to a clinically meaningful result for the patient. So, after the numbers have been crunched and the data analyzed, be sure to critically assess the results to make sure that a statistically significant outcome is meaningful and useful clinically. ■

Aron Shapiro is Vice President of Retina at Ora, Inc., in Andover, MA.

Dale Usner, PhD, is Vice President of Biostatistics at Statistics & Data Corporation in Tempe, AZ.