

THE ROLE OF BOOTSTRAP IN STUDY DESIGN

Julia Wang, The R. W. Johnson Pharmaceutical Research Institute; Rezaul Karim, Ortho-McNeil Pharmaceuticals, Inc;
Robert Medve, The R. W. Johnson Pharmaceutical Research Institute
Julia Wang, The R. W. Johnson Pharmaceutical Research Institute, Route 202, Raritan, NJ 08869-0602

Key words: bootstrap, data-based simulation, clinical trials, study design

Abstract: In pharmaceutical development, one element of an adequate and well-controlled efficacy study needed in conjunction with other studies to provide substantial evidence to support the claims of effectiveness of a drug for a particular disease or condition is the judicious selection of patients with the disease or condition being studied. A sufficient understanding of the effect and mechanism of action of the drug as well as the course of the disease is required to accomplish this task. Using various exploratory analyses, it may be possible to identify a patient population based on past studies that would clearly distinguish between the study drug and the control to reduce the study size and therefore the cost of drug development. However, if the studies are only carried out in this population, the use of the drug will have to be limited even though there are other patient populations that may also benefit from this drug, albeit to a lesser extent. We will present the use of bootstrap to guide the selection of patient population to find the best balance between the size of the study and the scope of the disease being treated.

1. Introduction

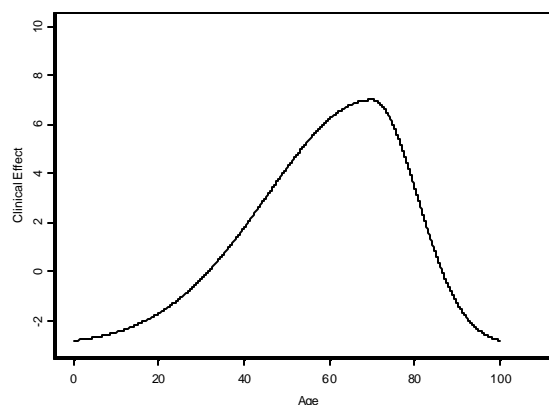
According to the Federal Food, Drug, and Cosmetic Act, pharmaceutical companies need to demonstrate the effectiveness of their drugs through the conduct of adequate and well-controlled studies in order to obtain marketing approval from the FDA. In general, the demonstration of effectiveness requires substantial and persuasive evidence from at least one and most often two adequate and well-controlled clinical studies [1]. The essential characteristics of such a study include clear statement of objectives, valid study design with appropriate controls, judicious subject selection, random treatment assignment, adequate blinding, well-defined and reliable assessments, and adequate analysis. In this paper, we focus on the aspect of judicious subject selection and discuss the use of bootstrap to help guide this part of the study design process.

2. The Indication and the Population

An active chemical entity interacts with the human body in certain ways. Due to the variations and differences in human bodies, this interaction will

produce no detectable clinical effects in some, undesirable effects in some, and hopefully, desirable effects in some. A chemical entity is effective in a certain sense if it is capable of producing a desirable effect in at least one human body. If a chemical entity affects a group of humans sharing a common characteristic in a desirable way, this common characteristic will be the indication for which the chemical entity will be efficacious in treating and this group of humans will be the target population for this chemical entity. In this case, the chemical entity will be referred to as a drug because the common characteristic is almost always a disease.

Figure 1: Clinical Effect vs. Age



Fully characterizing this target population is an evolving and iterative process. For ease of illustration, let's assume that there exists a one-dimensional numerically measurable clinical effect of a drug and this effect is correlated with a one-dimensional covariate such as age and graphing this effect against age produces a continuous curve (Figure 1). The drug produces a desirable effect for people between the age of 37 and 85. Running a clinical trial at the optimal age of 70 that produces the maximum effect on the curve (the mode) will require the least effort and cost to demonstrate the effectiveness of the drug. However, because of the continuity of the curve, all ages in a near neighborhood of the optimal age, say from 65 to 75, will also produce desirable effect, even though to a lesser extent. Including these ages will require a larger study size and therefore more costly to reach statistical significance but will result in a larger target population and less specific indication. In the following, we will refer to regions within the interval of ages from 37 to

85 in Figure 1 that produce the best trade-off between study size and target population size under a particular circumstance the ideal effect region for that circumstance.

In many cases, both the measurable clinical effect and the covariates are multi-dimensional and plotting the effect against the covariates will produce high-dimensional smooth planes. In addition, such planes can sometimes have multiple modes, meaning more than one possible indications and target population for the drug.

Normally, when a promising drug is ready to move into confirmatory human testing, some idea of its possible indication and target population have already been formulated based upon information gleaned from the chemical, non-clinical, pre-clinical, and earlier clinical research. In many cases, this realm of postulated effectiveness is less specific, close to or sometimes includes the ideal effect region.

When a confirmatory clinical study fails to reach any definitive conclusions with a non-significant p-value on the primary analysis, there can be many possible reasons for its failure. Flaws in either study conduct or design are often discovered and will be corrected if another similar study is to be carried out. However, frequently, the studies that are correctly designed and flawlessly executed will end up with statistically non-significant positive trends in the primary endpoint. This often indicates a lack of power resulted from insufficient sample size due to smaller than expected drug effect size (standardized with the variability estimate). Increasing the sample size in a similar subsequent study offers one solution but this can sometimes be too costly or unfeasible. In addition, statistically demonstrating a relatively small effect size with a large number of subjects doesn't put the drug in the best possible light.

Modifying the target population to include subjects enjoying bigger drug effect in the next study will be a good alternative with the following two scenarios. Figure 1 will be used again for illustration. In the first scenario, the study population included the mode and a portion of the optimal effect region as a subset, a restriction on age may be employed for the later studies. In the second scenario, the study population partially overlapped with the optimal effect region but missed the mode, upward or downward age adjustment should be used for the later studies.

However, restricting the study population will lead to a more specific indication limiting the scope of its potential uses. Rather than completely restrict the study

population, increasing the proportion of subjects with bigger drug effect while maintaining representation from subjects having smaller drug effect will be a better solution. The exact proportion can be estimated using bootstrap.

Since no drugs work on everyone, finding a target population for which the drug works should be a learning process. In an effort to gain regulatory approval for marketing, it's in the sponsor's best interest to showcase the drug in a most favorable setting. Contrary to the conventional logic employed to validate a principle, where you need to prove that it works under all conditions specified, to demonstrate the effectiveness of a drug, we only need to show that it works for some patient population with some disease. In the unfortunate case of several positive studies having been obtained only after a series of inconclusive ones, all part of an evolving and iterative process, the evidence provided by the positive studies should still be judged as persuasive and substantial.

3. An Example

A fixed-ratio combination analgesic was being developed for moderate to moderately severe pain. According to the guideline for the clinical evaluation of analgesic drugs, the superiority of the combination to each of its components (A and B) should be demonstrated. The first randomized, double-blind, parallel-group, active-controlled and placebo-controlled phase 2/3 trial to evaluate the efficacy and safety of this combination was conducted in subjects experiencing pain from an oral surgical procedure. Fifty subjects with moderate or severe baseline pain were enrolled for each treatment arm. Following the administration of study medication, subjects evaluated current pain relief from baseline and current pain intensity at 30 minutes, and 1, 2, 3, 4, 5, 6, 7, and 8 hours after medication. Pain relief was evaluated using a five-point categorical scale where higher scores indicate greater pain relief and pain intensity was evaluated using a four-point categorical scale ranging from 0=none, 1=mild, 2=moderate, and 3=severe. After the effect of the study analgesic wore off and subject took a supplemental analgesic, the pain relief scores were no longer collected. These missing scores were imputed using the last-observation-carried-forward method [2].

A number of variables were analyzed for this study, which included total pain relief (TOTPAR), sum of pain intensity difference (SPID), pain intensity difference (PID) at each observation point, pain relief (PAR) at each observation point, overall assessment of the medication, rate of remedication and time to remedication, ect. No primary variables were

preselected, as the FDA reviewers had preferred to examine the analgesic profile using all variables. TOTPAR, which was calculated for each subject by summing over the appropriate hourly pain relief evaluations, was the basis for sample size determination and a variable of high interest. In the following, we will be concentrating on this variable.

This study yielded significantly higher TOTPAR scores for the combination compared with component A or placebo ($p=0.0001$ for both). However, the comparison between the combination and component B didn't reach statistical significance ($p=0.3168$). Based on the observed results, a sample size of 465 per arm was needed for this comparison to reach statistical significance ($p<0.05$) with 80% power. However, it was further discovered that a significant interaction existed between the treatment and the baseline pain intensity ($p=0.0165$) and a sample size of 10 per arm was needed for this comparison to reach statistical significance with 80% power if we only enroll subjects with severe baseline pain intensity. For subjects with moderate baseline pain, the combination and component B are non-distinguishable. In this study, 19% of the subjects had severe baseline pain before study drug administration.

Table 1: Summary of TOTPAR by Baseline Pain Severity

TOTPAR	Combination	Component B
Moderate N Mean (SD)	37 10.23 (9.28)	38 11.78 (11.01)
Severe N Mean (SD)	13 18.31 (9.78)	12 6.21 (7.81)
All Subjects N Mean (SD)	50 12.33 (9.97)	50 10.44 (10.54)

Based on this result, it was further theorized based on the pharmacological properties of the combination and the component B that to increase the discriminate power of the pain model, more subjects with severe pain should be enrolled in later studies. However, how many subjects do we need per arm and what proportion of these subjects should have severe baseline pain?

For a given sample size per arm between 10 and 95, the proportions of subjects with severe baseline pain required to reach statistical significance with 80% power was estimated by bootstrap and presented in

Table 2. For example, for 50 subjects per arm, at least 48% of the subjects should have severe baseline pain.

Table 2: Proportion of Subjects with Severe Baseline Pain Needed to Reach 80% Power

Size per Arm	Proportion	Size per Arm	Proportion
10	95%	55	45%
15	83%	60	43%
20	73%	65	43%
25	63%	70	40%
30	58%	75	40%
35	55%	80	40%
40	53%	85	40%
45	50%	90	38%
50	48%	95	38%

For each combination of a sample size and a proportion of severe baseline pain, the power was simulated using bootstrap. For example, for 50 subjects per arm and 30% severe pain, 15 and 35 subjects were re-sampled respectively with replacement from the group of subjects with severe baseline pain and the group of subjects with moderate baseline pain within that treatment arm. This re-sampling process was repeated 100 times and the bootstrap power and mean p-values were estimated from these 100 runs. This process was repeated for sample sizes per arm ranging from 10 to 95 in increments of 5 and for proportions of severe baseline pain ranging from 5% to 95% in increments of 5%.

Figure 2 presents the four-panel rotation contour plots of the smoothed mean bootstrap p-values against the sample size and the proportion of severe subjects.

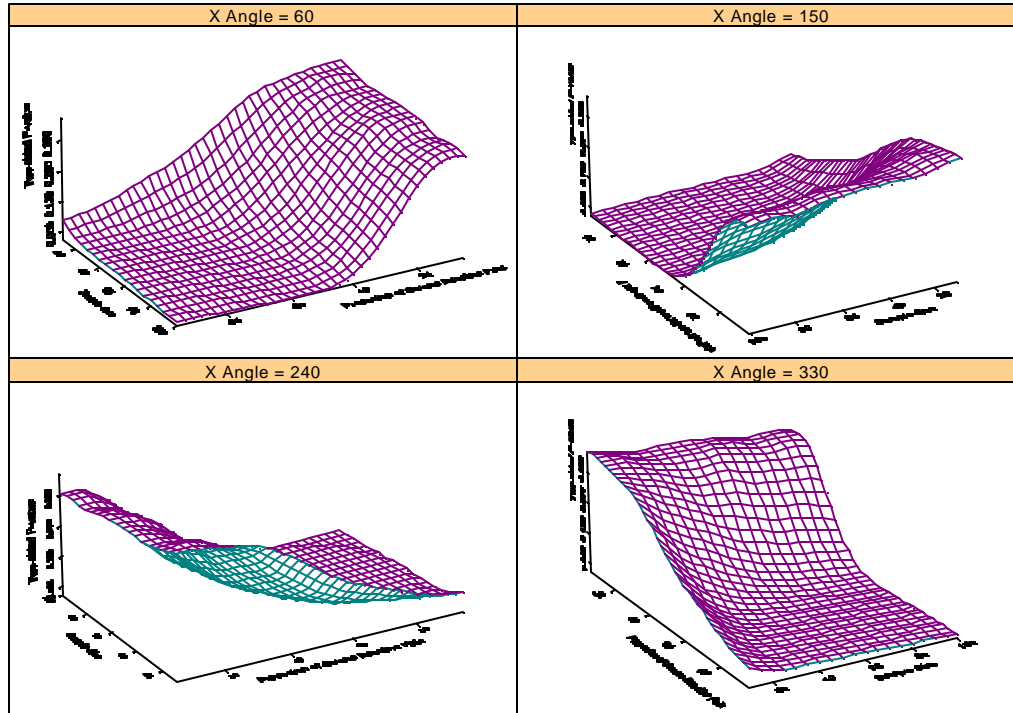
Three more similar dental pain studies were carried out. Each of these three studies enrolled 80 subjects per arm. Subjects were still required to have moderate or severe pain as a result of an oral surgical procedure. In addition, unlike the first study in which the oral surgical procedure were required to involve at least one partial boney impaction, they required to involve extraction of two or more impacted third molars requiring bone removal and if only two impacted third molars were extracted, they should be ipsilateral. A few external experts helped us to identify this population of subjects because from their experience, a much higher percentage of this population would experience severe baseline pain.

All three new studies succeeded in statistically separating the combination analgesic from both of the components. Over 30% of the subjects in each study

experienced severe baseline pain. It's highly likely that the improved discriminate ability of the dental pain model [4] is mainly driven by the severity of the oral procedure performed and the measurable baseline pain is partially correlated with this procedure severity. By trying to enroll subjects undergoing more severe oral procedures in the hope of increasing the proportion of

subjects having more severe baseline pain, the right study population was serendipitously found in all three studies. Taken as a whole, the probability of having these three studies positive simultaneously is 0.936×10^{-6} using TOTPAR, under the null hypothesis of no difference between the combination and component B.

Figure 2: Mean simulated P-values vs. Sample Size per Arm And Proportion of Subjects with Severe Baseline Pain



4. The Explanation and Generalization

During the process of gaining regulatory approval for the marketing of a potential new drug, the uncertainties inherent in the clinical trials are in a way characterized by p-values. A p-value is a random variable with a uniform distribution under the null hypothesis. Its distribution under the alternative hypothesis determines the power of the study and this power is the factor that we seek to optimize. However, the alternative distribution is under the influence of many known or unknown covariables and can only be estimated if the alternative hypothesis is repeatedly sampled under the same condition, a costly and impossible maneuver.

Bootstrap [5] is a resampling method used under the principle that resampling the sample approximates resampling the population. With bootstrap, the current study population is treated as the population from which future study subjects would be enrolled. By

randomly drawing these subjects out with replacement to form simulated studies repeatedly, various statistical quantities can be estimated.

For example, the distribution of a p-value under the alternative hypothesis can be estimated as follows: resample the treatment groups separately, combine the new samples and calculate the p-value, then repeat this process many (generally over 1000 if no resource constraint) times.

From this estimated distribution of the p-values under the alternative hypothesis, the power of any future studies, designed, executed, and analyzed in a similar way as the existing study on which the bootstrap is based, can be estimated by the proportion of simulated studies producing significant p-values.

Currently, projecting the power of future studies is normally done using the point estimates obtained from

the existing studies along with certain parametric distribution assumptions. Bootstrap offers a better alternative with improved asymptotic properties. In addition, it provides a means of estimating the power of future studies even when the study population and/or method of analysis are modified to be different from that of the existing studies.

In the regular clinical trial setting, rather than obtaining a random sample from a population, we get a convenience sample. In order to use the mainstream statistical theory, this sample is assumed to be randomly sampled from a super-population post-fitted to have the characteristics exhibited by the observed sample. These characteristics are somewhat controlled by the inclusion and exclusion criteria of the protocol. When we start a new study using the same inclusion and exclusion criteria as that used in the observed study, hoping to capture the same sort of patients, we can treat this new study as a random sample from the same super-population and use bootstrap to estimate its behavior.

If we change the inclusion and exclusion criteria, the new study, when carried out, can only be considered as a random sample from a different super-population post-fitted to have the new characteristics. However, we can still use our observed study to estimate the behavior of this new study, if we can modify the sampling mechanism in such a way that the simulated studies can be treated as simple random samples from the new study.

Suppose that after a clinical study is completed, the study population can be partitioned into several subgroups according to the efficacy responses. The responses within each subgroup are relatively homogeneous while those across subgroups vary noticeably. These subgroups don't have to be the pre-stratified subgroups. For ease of illustration, let's assume there are three such strata with $p_1\%$, $p_2\%$, and $p_3\%$ each, and these three strata have characteristics A, B, and C, respectively. The super-population post-fitted for this study should have $p_1\%$ of A, $p_2\%$ of B, and $p_3\%$ of C. Since these three strata produce different magnitude of efficacy responses, we may want to identify an optimal ratio of these three strata for the next study while maintaining representation from all three strata. Bootstrap can be used to simulate the study for any number of subjects and for any combinations of non-zero $q_1\%$, $q_2\%$, and $q_3\%$ as long as $q_1+q_2+q_3=100$. These simulated studies will point to new super-populations.

For example let's assume the sample size per treatment arm is 100. We will randomly sample q_1 subjects with

replacement from each treatment group in the first stratum, q_2 subjects with replacement from each treatment group in the second stratum, and q_3 subjects with replacement from each treatment group in the third stratum. Then the three samples are combined to form one simulated study. This is equivalent to getting a random sample from the super-population with $q_1\%$ of A, $q_2\%$ of B, and $q_3\%$ of C. Various distributions of interest can be estimated for this new super-population by repeating the above process many times. In addition, these estimated distributions converge in distribution with a second-order rate to the underlying distribution of the super-population.

A similar process can be utilized for a variety of population-modifying problems, no matter how complicated the situation is.

5. Conclusions

In pharmaceutical development, data-based simulation such as bootstrap can be very useful to guide the design of future studies based on the studies already carried out. This effort should especially be made after phase 2 studies are completed and the confirmatory phase 3 programs are being planned. In the phase 3 stage, if sufficient time exists between the completion of consecutive studies to allow the modification of the ongoing or to-be-initiated studies based on completed studies, bootstrap should also be done. Conducting clinical trials is a very time-consuming and costly business and effort should be made to extract as much useful information as possible from the data already collected and bootstrap can help us accomplish this goal.

6. Reference

1. FDA Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products, May 1998
2. Wang, Julia: The Applications of Multiple Imputation Using a Commercial Software in Clinical Trials with Dropouts, Proceedings of Biometrics Section, 53-60, JSM 1999
3. FDA Guideline for the Clinical Evaluation of Analgesic Drugs, Dec. 1992.
4. Averbuch, M. and Katzper, M: Baseline Pain and Response to Analgesic Medications in the Postsurgery Dental Pain Model, J Clin Pharmacol, 40, 133-137, 2000.
5. Efron, Bradley and Tibshirani, Robert: An Introduction to the Bootstrap, Chapman & Hall, 1993.