A profile-based stratified randomization and its application to a double-blind, placebo-controlled clinical trial

Andrew Magyar, Cornelia Haag-Molkenteller, Jihao Zhou[1]

Allergan, PLC, Dublin, Ireland

Abstract

Stratified randomization is a common technique used in clinical trials to control for important baseline characteristics. However, as the number of stratification factors increases, the number of strata will become too large to handle in practice. A double-blinded, placebo-controlled, phase 3 trial was designed in which there were four factors that needed to be accounted for using stratification. To determine strata for randomization, a profile-based stratification method was proposed which used each subject's baseline profile to calculate their individual propensity score. The parameters in the propensity score function were estimated based on patient data from two similar previous phase 3 clinical trials. This method led to less variation in the estimate of the *p*-value while still providing an unbiased estimate.

Key Words: propensity score, randomization, stratification, clinical trial, experimental design

1. Introduction

In clinical trials, it is often necessary to account for demographic and baseline disease characteristics that could confound the effect of the treatments under evaluation. To the previous goal, it is desired to have treatment arms where the distribution of possible confounding factors is the same within each arm. This characteristic is referred to as having balanced treatment arms - that is balanced with respect to the distribution of the confounding factors. By having balanced arms, the effects that the confounding factors have on the endpoint of interest is expected to be the same within the treatment arms. Consequently, any differences observed may be attributable to something other than the confounding factors, hopefully, the clinical effect of the treatments. Note that being balanced does not necessitate the more ambitious goal of having the distributions within the treatment arms be reflective of the distribution of the factors within the clinical population of interest.

In order to obtain treatment arms that are balanced one method is to employ stratified randomization. Stratified randomization entails splitting enrolled patients into separate groups (strata) based on the confounding factors of interest and then randomizing them to the treatment arms within each stratum. When there are several confounding categorical factors to consider, each with multiple levels, then it is necessary

[1] Corresponding Author: Jihao Zhou, PhD
2525 Dupont Drive, T2 6F, Irvine, CA 92612, USA
e-mail: zhou_jihao@allergan.com

to define a stratum for every possible combination of factors and levels in order to achieve balanced treatment arms. Unfortunately, when the number of confounding factors and levels is large, implementing stratified randomization is difficult, if not impossible, in practice since some factor-level combinations may be underrepresented, or not represented at all, because of constraints on sample sizes.

It is important to note that achieving balanced treatment arms is not the goal in and of itself. The goal is to balance the potential 'effect' that the confounding factors have on the endpoint of interest. Balanced treatment arms is merely one method to achieve this goal. With balanced treatment arms, in theory the effect of each confounding factor individually (as well as any interactions) are cancelled out on a factor by factor (and interaction by interaction) basis, thus implying a cancellation of the aggregate effect. However, for the purposes of assessing a treatment effect one is only concerned with adjusting for the aggregate effect of several confounding factors, the fact that the individual effects of each confounding factor and interaction effects are equal between treatment arms adds nothing to the conclusion. To further illustrate the previous concept, consider a hypothetical placebo-controlled clinical trial testing the efficacy of an experimental new hypertensive. It is known that higher blood pressure is associated with older age and being overweight. Suppose there is an imbalance in the distribution of age between the placebo and active arms, e.g., there are more young people in the active arm. The effect of this imbalance in age could theoretically be counteracted by an imbalance in the distribution of weights between the arms. That is, if the active arm tends to have younger patients who are heavier and the placebo arm has older patients who tend to be slim, then the net effect of these two factors on blood pressure may be equal between these two arms despite the fact the age and weight are far from balanced.

In this paper, an alternative method is presented to control for confounding factors based on the concept of propensity-score (Rosenbaum PR and Rubin DB 1983). Much like the example in the previous paragraph, the proposed method circumvents the need to establish balanced treatment arms, but rather obtains arms where the aggregate effect of the confounding factors & interactions is equal between treatment arms. This method was applied to a recently concluded double-blind, placebo-controlled clinical trial testing the efficacy of an investigational compound. Simulations based on the results of the trial were performed and the findings presented.

2. The Profile-based Stratification Approach

The study of interest was a multicenter, double-blind, randomized, placebo-controlled, parallel-group, phase 3 clinical trial. The primary endpoint was the change from baseline in the primary efficacy variable. It was needed to control for four factors known to be confounders with the primary endpoint of interest. Three of these factors were dichotomous and the last was numerical. Patients were to be assigned in a 1:1 ratio into the active and control arms. Given there were four factors to be accounted for, stratified randomization was not feasible for this study. Instead, a novel stratification approach based on the concept of propensity score was implemented with the intent of balancing the effect of the four confounding factors.

Rosenbaum and Rubin (1983) introduced the concept of propensity score for a subject as the conditional probability of assignment to a particular treatment versus

control given a vector of observed covariates. Consider the event a subject achieves at least a 50% reduction from baseline in the endpoint of interest given that the patient is in the active arm (coded as a Bernoulli variable, 1 being success). Let p, denote the probability that this event happens. This probability could depend on the values of the confounding factors, let

$$p\left(\underset{\sim}{x}\right) = \Pr\left(Y = 1 \middle| \underset{\sim}{X} = \underset{\sim}{x}\right)$$

Further assume that this follows a logistic regression model (using the logit function as the link function),

$$p\left(\underset{\sim}{x}_i\right) = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id}\right)}$$

where

- p_i: the i^{th} patient's estimated predictive probability to achieve at least a 50% reduction from baseline in the endpoint of interest given that the patient is in the active arm
- x_{ij}: the i^{th} patient's value for the j^{th} confounding factor
- β_j: the coefficient corresponding to the j^{th} confounding factor

In order to estimate the parameters of the logistic regression model, the model was fit via maximum likelihood estimation using data from two previous pivotal phase 3 studies.

Using the patients from the two pivotal phase 3 trials, the median value of the propensity scores was calculated to be 0.453. This value is then used to define strata for stratified randomization for the study of interest. If $p_i \geq 0.453$, the i^{th} patient will be assigned to stratum A; otherwise, the patient will be assigned to stratum B. Patients will then be randomized in a 1:1 ratio to either the active or placebo arms.

3 Simulations

Recall that the threshold of 0.453 corresponded to the median propensity score of patients from the two previous pivotal phase 3 trials which were used to estimate the parameters of the logistic regression equation. As to be expected, the median propensity score for the patients randomized in the study of interest need not be the same, thus leading to an imbalance between the number of patients in stratum A and stratum B. Roughly 20% of the patients were in stratum A whereas 80% were in stratum B. Nonetheless, the study did achieve its endpoint of interest, the difference between active treatment and placebo was found to be statistically significant.

In order to address the merits of the proposed propensity score-based stratification method, a simulation experiment was performed utilizing the data from the clinical study. Boot-strapped samples of N subjects (sampled with replacement) from each treatment arm were obtained using two methods:

I) No stratification - a random sample of N subjects from each treatment arm was taken.

II) Propensity score stratification - within each treatment arm $0.2\times N$ patients were sampled from stratum A and $0.8\times N$ patients were sampled from

stratum *B*. When $0.2\times N$ or $0.8\times N$ did not yield integers, the value was rounded to the nearest integer.

For each simulation run, the change from baseline in the endpoint of interest was taken for each patient in the sample and a two-sided, two-sample *t*-test was performed using treatment arm as the factor and the *p*-value was recorded. The total sample size, *N*, was varied from 5 to 40 subjects. For each sample size, the simulation was run 10,000 times.

4. Results

For a fixed sample size, the following sample statistics were taken of the *p*-values from the 10,000 simulation runs:

i) Average *p*-value – for each sample size, the sample average of the 10,000 *p*-values was taken.
ii) Power – for each sample size, the proportion of *p*-values that were less than 0.05 was obtained (i.e. the number of simulation runs that rejected the null hypothesis).
iii) *p*-value variance – for each sample size, the sample variance of the 10,000 *p*-values was calculated.

The figure below is a scatterplot of the sample size in the horizontal axis and the average *p*-value in the vertical axis. The values obtained from using no stratification are in red whereas those using the propensity score stratification are in black.

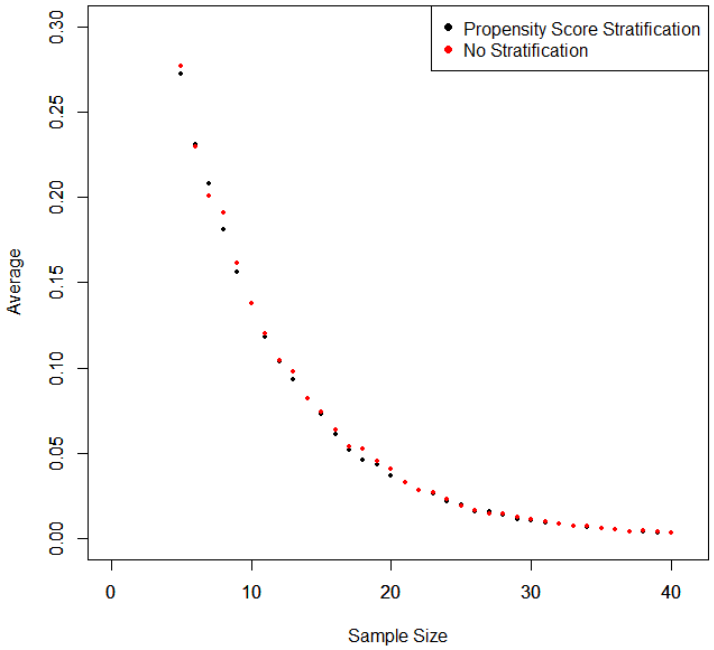

Figure 1. Average *p* -values by sample size

As is evident, whether one uses no stratification or propensity score stratification, the average *p*-value is the same.

What is more of interest is whether one method is arriving at the correct conclusion more than the other (that is to reject the null hypothesis of no treatment difference). To this end, the power was plotted versus the sample size in the scatter plot below.

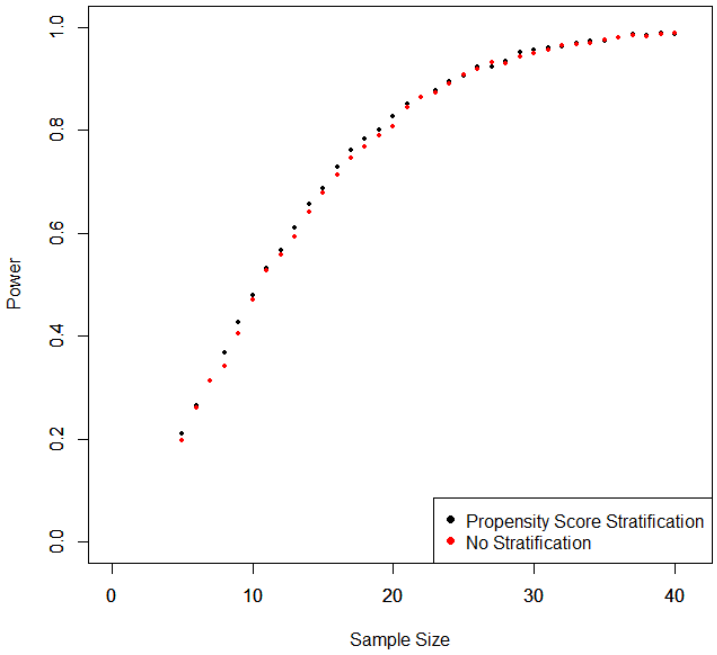

Figure 2. Empirical power by sample size

What is apparent is that both methods are rejecting the null-hypothesis the same proportion of time (i.e. have the same power).

While the previous results do not seem to provide much merit to the propensity score stratification method, they are not surprising. Recall that 'on average' random sampling with 1:1 randomization will obtain samples in each arm that are not only balanced with respect to the confounding factors, but such that the distributions of the confounding factors in each treatment arm is the same as the underlying population. For example, if 20% of the population is left-handed, then random sampling with 1:1 randomization will 'on average' provide samples such that 20% of patients in each treatment arm are left-handed. Consequently, ‘on average’ the results of a *t*-test will be unbiased and lead to the correct conclusion to reject the null hypothesis.

However, ‘on average’ does not guarantee the confounding factors will be balanced for a single sample, such as in a clinical study. This previous fact is the

motivating factor for the development of methods that account for the previous short-coming of simple random sampling such as stratified randomization or propensity score-based randomization. Plotted below is scatter plot of the variance of the *p*-value versus the sample size.

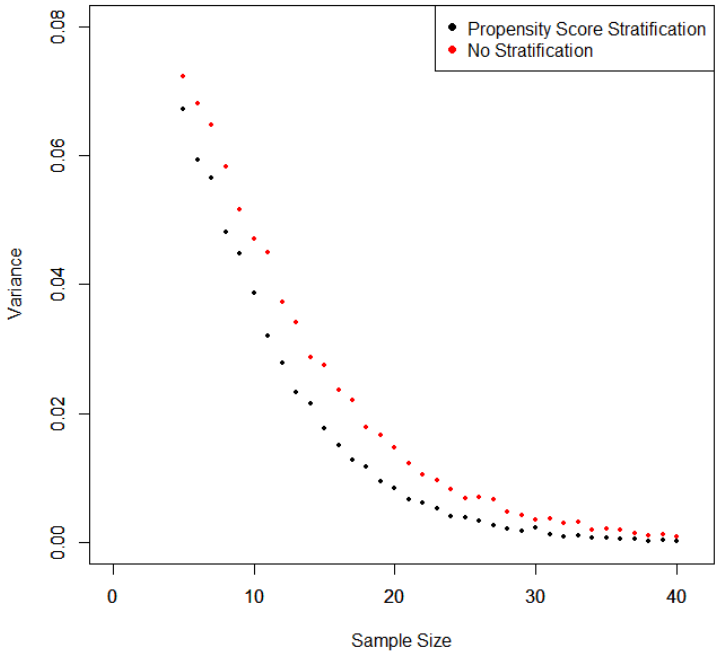

Figure 3. Variance of *p*-values by sample size

For every sample size, the variance of the *p*-value using propensity score stratification is less than not using stratification at all.

5. Discussion

As demonstrated in the last section, the statistical merit of the propensity score stratification is in that it provides a more reliable (less variable) estimate of the *p*-value while still maintaining an unbiased estimate of the *p*-value with the same power as if one had not stratified. Perhaps more of interest is its logistical merit in that it provides a means to control for the effect of multiple confounding factors without having to define a large number of strata for randomization.

Unlike standard stratified sampling, propensity score stratification will not guarantee that the distribution of the confounding factors in both treatment arms is balanced. Instead, propensity score stratification obtains treatment arms such that the 'aggregate effect' that the confounding factors have on the endpoint of interest is 'balanced' between treatment arms. Since both treatment arms are balanced in a way such that the aggregate effect of the confounding factors is the same, the variation in the endpoint of interest caused by the confounding factors is adjusted

for, thus leading to less variation in the *p*-value. In essence the aggregate effect of the confounding factors is able to be reduced down to a single metric. Consequently, one can think of propensity score stratification as a dimension reduction method.

Acknowledgements

Authors would like to acknowledge Allergan, Inc. for the support of using innovation approach to clinical trial practice.

References

D'Agostino, R (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 17: 2265-2281.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41-55.

Rubin DB (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials Statistics in Medicine **26**:20–36.