PharmaSUG 2014 - Paper SP08

Methodology for Non-Randomized Clinical Trials: Propensity Score Analysis

Dan Conroy, Ph.D., inVentiv Health, Burlington, MA

ABSTRACT

Randomized clinical trials serve as the gold standard for all research trials conducted within the pharmaceutical industry. When trials are not properly randomized, there is a potential for bias in all subsequent statistical analyses. When proper randomizations are not in place for a trial, methods exist that can help researchers draw valid conclusions. This paper summarizes and demonstrates some potential methods that can be used in this scenario. In particular, propensity score methodologies are presented and discussed with illustrative examples and SAS® code. The examples and issues discussed herein were selected with the intent that researchers may find them informative, relevant, and applicable in a variety of scenarios, so that they may mimic and apply these methods in their own statistical analyses. The content is directed at users of SAS software with an intermediate understanding of standard statistical concepts and methodologies.

INTRODUCTION

Researchers are well aware that properly randomized clinical trials are the basis for the majority of clinical trials conducted within the pharmaceutical industry. The same can be said for trials within the biotechnology industry as well as the medical device industry. Trained statisticians and biostatisticians are aware that conducting trials without a proper randomization in place allows the potential for biases in subsequent statistical analysis. This increases the potential of drawing incorrect conclusions.

Typically, it is a straightforward matter for investigators to ensure a randomization is in place for trials in Phases I-III, because these trials tend to be prospective in nature. However, the same is not always the case for post-marketing trials in Phase IV work. These trials tend to be observational in nature, so one cannot always guarantee that subjects or experimental units received treatments in accordance with a correct randomization scheme.

The importance of Phase IV trials is recognized due to their typically large sample sizes and application of treatments within a real world setting. Thus, testing treatments within these trials may be considered an ideal environment to further confirm conclusions drawn in earlier phases. However, statisticians must recognize that since no randomization may be in place, their methodologies need to account for external confounding or lurking variables in order to draw unbiased and correct conclusions.

One set of methodologies that can be used in this scenario involves first calculating propensity scores as an early step in the statistical analysis and then incorporating these quantities in subsequent portions of the analysis. Calculation of propensity scores is typically done with an application of the LOGISTIC procedure in SAS. Propensity scores can be outputted using options within PROC LOGISTIC. These propensity score quantities can be incorporated into later models or hypothesis testing as quantitative or categorical variables.

The concepts and ideas suggested in this introductory section will be further discussed as this paper progresses. They will be utilized and employed within the context of several examples in order to fully illustrate the general idea as well as some of the nuances that may arise. Select SAS code will be presented as needed with the intent and hope that the reader may employ these methods without significant difficulty.

PROPENSITY SCORE DEFINITION

In this section, we will first define what a propensity score is. Once this concept is established, we will demonstrate how they can be calculated and outputted by appealing to an application of PROC LOGISTIC in SAS.

A propensity is a noun typically used to refer to a tendency to do something. This is the working definition within Standard English, and the same concept will apply within the context of our statistical methodology. We believe that the concept can be best defined by way of example, as will be done next.

Let us suppose that we have an observational trial which is completely retrospective in nature. The subjects in our data were either treated or not with a particular treatment, but not according to any randomization. Instead, let us suppose that physicians who examined the subjects either prescribed the treatment or not. Furthermore, we have various other data available for each subject, including gender, age, race, BMI, height, weight, blood pressure, and perhaps others.

Thus, it is indeed known whether or not each subject in our trial received the treatment. For our purposes, we will be interested in the probability that each subject received the treatment. Here lies the crux of the matter. We want to calculate the probability of treatment using our other available data (such as gender, race, BMI, height, weight, blood pressure, and so forth) as covariates. The propensity score for each subject can be defined as the probability that a subject was treated or not.

Most statisticians are familiar with logit models, and are aware that these models look at the odds of an event. Odds are known to be simple functions of probabilities, defined as:

Odds of an Event = Probability(Event Occurs) / Probability(Event doesn't occur)

Thus, if the odds of an event are known, one can easily solve for the probability that the event occurred since odds are a simple function of probabilities. This probability is our propensity score, and it can be easily outputted using PROC LOGISITIC without the user having to do any manual calculations. Let us suppose that we have treatment, gender, race, BMI, height, weight, and blood pressure (BP) recorded in a data set named FINAL. Propensity scores can be outputted into a data set named SCORES with variable name PS as follows:

```
proc logistic data=final;
Class treatment gender race;
Model treatment = gender race BMI height weight BP;
Output=SCORES pred=PS;
run;
```

Thus, the output data set SCORES contains the same information as the input data set, FINAL, as well as the propensity score for each subject in a variable called PS. Remember, that the propensity scores are the probability that each subject was actually treated, modeled as a function of each subject's gender, race, BMI, height, weight, and blood pressure.

BALANCE ASSESSMENT TABLES

In this section, we follow up the previous propensity score definition with a concrete example of the ideas and concepts previously presented. We again suppose that we have observational data with subjects who were treated or untreated, as well as their gender, age, race, BMI, height, weight, and blood pressure. Let us further suppose that the treatment is intended to cure some particular disease, and we are interested in the primary outcome of whether or not subjects died by the end of the study. If one were to analyze the data directly, there is a possibility for bias because there was no randomization to specify which subjects were treated and which were not. For example, it may be the case that all the younger patients in the study were treated, and all the older patients in the study were not. If one analyzes the data directly in this artificial example, it may appear that subjects were less likely to die if they were treated and more likely to die if they were untreated. However, this is a very rudimentary approach because our intuition tells us that the older patients (who were untreated) are perhaps more likely to die anyway, regardless of whether or not they were treated. Thus, age may be a lurking variable that biases our conclusions if we don't attempt to account for this possibility. Propensity scores are used to attempt to account for the biases that may be inherent as a result of the lack of randomization. This point is stressed as it is the foundation of all the ideas that we are using in this paper.

To calculate propensity scores, we first fit the logit model with treatment as the response and the remaining information as the covariates. The model is:

Logit(odds(treated)) = Terms involving the available covariates

It is noted that the covariates on the right hand side of the model equation could of course include linear terms, interaction terms, and so forth. After outputting the propensity scores as explained in the previous section, one then has a data set that includes a propensity score for each subject.

The idea for further analysis will then be to use the propensity scores in a way that may account for lurking variables, such as age was in the discussion above. We need to do something to assess whether or not our propensity scores are adequately accounting for any other potential lurking variables. One way to assess the situation is to create a

balance assessment table. This table often includes three columns. The first column includes each covariate that might be a lurking variable. The second column includes a *p value* to assess the relationship between each potential lurking variable and the treatment. The third column includes a *p value* to assess the relationship between each potential lurking variable and the treatment in the presence of our previously calculated propensity scores. We present a balance assessment table in Table 1, and describe its features and interpretation after.

Covariate	**P-value from Model: Covariate=Treatment**	**P-value from Model: Covariate = Treatment + PS Quintile**
Gender	0.02	0.86
Race	0.03	0.67
Age	0.04	0.03
More variables…	…	…

Table 1. Balance Assessment Table

In Table 1, the *p values* in the second column may be found by modeling the covariate from the first column against treatment. In the case of a continuous variable such as age, one might fit an ANOVA model with age as the response and treatment as the predictor. In the case of a categorical variable such as gender, one might fit a logistic regression model with gender as the response and treatment as the predictor. The *p values* in the third column may be calculated similarly. First, we note that the propensity score quintiles are used as the predictors in the third column. Quintiles are used because there are places in the literature that suggest doing so for various reasons. Thus, one could calculate the *p values* for a continuous covariate such as age by fitting an ANOVA model with age as the response, and treatment and propensity score quintiles as the predictors. In the case of a categorical variable such as gender, one could fit a logistic regression model with gender as the response and treatment and propensity score quintiles as the predictors.

The purpose of creating a balance assessment table as show in Table 1 is to determine whether or not our propensity scores might help to offset potential biases introduced due to the non-randomized nature of our observational data. Interpreting the information found in Table 1 will be discussed next.

The first covariate in the table is gender. One notices that the *p value* for the gender=treatment model in the second column is 0.02 and the *p value* for the gender=treatment + propensity score quintile model in the third column is 0.86. The *p value* in the second column indicates that there is indeed a significant difference in gender between the two treatments. This means that the treatment was not assigned in equal proportions to the two genders in our observational study. This indeed may cause bias in our conclusions if one did nothing to account for it. The *p value* in the third column is 0.86 and indicates that there are no longer statistically significant differences in gender between the treated and untreated groups when propensity score quintiles are incorporated in the model. Thus, one may conclude that incorporating propensity scores in subsequent analysis for the primary outcome measure may help account for differences in gender between the treated and untreated groups.

An analogous interpretation is found for the race covariate. The *p value* in the second column is 0.03 and that in the third column is 0.67. This would indicate that there are significant differences between races between the two treatment groups but these differences are no longer significant after adjusting for propensity scores. Thus, incorporating propensity scores in subsequent analyses may indeed help account for the differences in race between the treated and untreated groups.

The situation found for the age covariate is different. Both the *p values* in the second and third columns are less than 0.05. This indicates that there are differences in age between the treatment group before *and* after adjusting for propensity scores. Thus, incorporating propensity scores in subsequent analyses may not do enough to prevent biasing our conclusions. One will have to carefully attempt to account for this fact when analyzing the primary outcome variable.

USING PROPENSITY SCORES TO ANALYZE THE PRIMARY OUTCOME

The discussion found thus far in this paper is focused primarily on establishing the prerequisite definitions, concepts, and ideas that are needed to properly conduct a final analysis for the primary outcome variable. Depending on the nature of the available data and study design, the propensity scores may be used in various ways to conduct the proper and required final analysis. In our example case, the reader is reminded that we are supposing that we have one primary outcome of death, and we are interested in whether or not it is related to treatment. The discussion found in this section will illustrate several ways in which the relationship between death and treatment might be assessed.

Initially, one might take the rudimentary approach of fitting a logistic regression model with death as the response and simply treatment as the single covariate. Let us suppose that this procedure is done, and that we find the *p value* for treatment is 0.02 in the resulting model. As discussed throughout this paper, this approach does not account for the fact that our data is observational in nature, and no randomization is in place. Thus, it may appear that treatment is a significant predictor of death, but it may be that other lurking variables are present that are biasing this simplistic conclusion.

A better approach is to fit the logistic regression model with death as the response, and treatment and propensity score quintiles as the predictors. Once this is done, let us suppose that the *p value* for treatment 0.64, indicating that treatment is no longer significant after adjusting for propensity score quintiles. We now see that there does not appear to be significant relationship between treatment and death. As the reader recalls, we found in Table 1 that adjusting for propensity scores appeared to help balance out differences in gender, and race, but this was not the case for the differences in age. Thus, one would likely want to fit a final logistic model with death as the response, and with treatment, age, and propensity score quintiles as the predictors. This model may likely be our best model to use as it should do the best job balancing out the differences in all covariates. Final conclusions regarding the significance of treatment should be drawn from this model.

In some situations, it may be preferable to analyze the data strictly using a categorical approach. The first analysis that one may conduct is the well-known chi-square test for independence using a 2x2 contingency table. The table would simply contain counts for deaths (yes or no) and treatments (treated or untreated). Suppose the independence test is conducted and then a *p value* of 0.04 is found. If one draws conclusions from this approach, it is found that death and treatment are not independent. However, this approach fails to account for any lurking variables that may be present as a result of the observational nature of our data.

A more appropriate categorical analysis may be conducted as follows. One could again derive the quintiles of the propensity scores. Then, the Breslow-Day test could be used to test for an association between the primary outcome of death and the treatment, while stratifying on the propensity scores. This approach helps to account for treatment imbalances. If a large *p value* is found, one may conclude that there is no evidence that death and treatment are related, after adjusting for propensity scores.

POTENTIAL LIMITATIONS

In this section, we will present some of the limitations of the methods discussed in this paper.

Throughout several examples in this paper, the approach was to split the propensity scores into quintiles. This approach has been suggested in the literature. However, one possible limitation may be that the sample size within each quintile may be small if our overall sample size isn't very large.

Another important limitation is that this approach may not be able to account for all lurking variables. In any observational trial, there may be lurking variables that we are not aware of. Using propensity scores analysis methods may be helpful, but one can usually not be certain that we are accounting for all biases due to variables that we are not aware of.

CONCLUSION

In conclusion, we note that propensity score methodologies offer statisticians a way to possibly eliminate bias in observational data. The examples and methods presented here offer one way in which propensity scores may be used for this purpose, but it is noted that they may be used in other ways as well. One additional point that is worthy of emphasis is that we can never be certain that there are no other lurking variables in observational data, so it is suggested that statisticians draw their conclusions with some caution whenever they are analyzing non-randomized data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Dan Conroy, Ph.D.
Enterprise: inVentiv Health
Address: 1 Van de Graaff Drive
City, State ZIP: Burlington, MA 01803
E-mail: Daniel.conroy@inventivhealth.com
Web: http://www.inVentivHealthclinical.com