Paper 4687-2020

A Doctor's Dilemma: How Propensity Scores Can Help Control for Selection Bias in Medical Education

Deanna Schreiber-Gregory, Henry M Jackson Foundation for the Advancement of Military Medicine

ABSTRACT

An important strength of observational studies is the ability to estimate a key behavior's or treatment's effect on a specific health outcome. This is a crucial strength as most health outcomes research studies are unable to use experimental designs due to ethical and other constraints. With this in mind, one drawback of observational studies (that experimental studies naturally control for) is that they lack the ability to randomize their participants into treatment groups. This can result in the unwanted inclusion of a selection bias. One way to adjust for a selection bias is by using a propensity score analysis. In this paper, we explore an example of how to use these types of analyses. In order to demonstrate this technique, we seek to explore whether clerkship order has an effect on National Board of Medical Examiners (NBME) and United States Medical Licensing Examination (USMLE) exam scores for 3rd year military medical students. In order to conduct this analysis, a selection bias was identified and adjustment was sought through three common forms of propensity scoring: stratification, matching, and regression adjustment. Each form is separately conducted, reviewed, and assessed as to its effectiveness in improving the model. Data for this study was intended to imitate data gathered between 2014 and 2019 from students attending Uniformed Services University of Health Sciences (USUHS). This presentation is designed for any level of statistician, SAS® programmer, or data scientist or analyst with an interest in controlling for selection bias.

INTRODUCTION

The purpose of this study was to explore the possibility of a connection between the sequence one chooses for their clerkship year of medical school and the resulting grades they obtain on major examinations, specifically USMLE Step1 and Step2CK. A secondary, and arguably equally important, goal of this study was to explore the possibilities and complications of utilizing a propensity score analysis to control for selection bias in an observational educational dataset.

Traditionally, randomized control trials have been the standard research design when estimating causal treatment effects. The main advantage to these types of studies lie in the fact that researchers are able to randomly distribute participants into treatment groups, thus allowing them to reduce selection bias and derive causal inferences from the resulting analyses. Nevertheless, these types of studies are not always feasible due to small sample sizes, budgetary constraints, and ethical limitations, and are often restricted to subpopulations that end up limiting the generalizability of results (Rubin, 2007). Observational studies, on the other hand, have the ability to evaluate treatment effectiveness in a home or healthcare environment which helps increase generalizability and decrease concerns about sample size, budgetary impact, and ethical boundaries; however, true randomization of participants within these studies are near impossible to obtain. Without

randomization, the differences in baseline covariate distributions between treated and untreated participants confound the comparison of outcomes between treatment groups. This eliminates the methodological support of causal inferences generated from these types of studies. Nevertheless, observational studies are still needed and continue to be routinely implemented with the goal of estimating causal effects for a variety of treatment outcomes; therefore, a series of alternative pseudo-randomization techniques have been developed as viable alternatives to true randomization in order to explore causal inferences and attempt to earn back some of the methodological support lost in the transition from randomized control to observational health studies (Ross, et al., 2015). One such technique for pseudo-randomization is through use of propensity scores. Through this technique, exposure is modeled within a preliminary structure based on the investigators' assumptions and understanding of the sampled dataset. This preliminary model outputs a probability of exposure - the propensity score - which is then included in the response model. Propensity scores have the ability to take on the form of a covariate, can be categorized into subclasses for stratification, can be transformed into weights for standardization, or can be used in a matching analysis. Each of these methods has the same goal of confounder balancing between the exposure groups in order to reduce selection bias. The choice in method is dependent on the nature of the question, the size of the dataset, the number of possible confounders, and the prevalence of exposure and outcome (Ross, 2015).

Since the ground-breaking introductory paper by Rosenbaum and Rubin in 1983, propensity score analyses have increasingly been utilized in a variety of different fields, including pharmaceutical medicine, health, education, and economics. All of these fields have variables with which one could infer both causal and correlational relationships depending on the complexity and structure of the interactions involved. Given this observation, it is worthy to note that not all models are created with the appropriate assumptions and structure for implied causal relationships, even with propensity score utilization. For example, in one study that utilized a secondary propensity score analyses of electronic health data to explore the effects of a Medicare Part D prescription drug program for individuals with serious mental illness (Stuart et al., 2013), the researching statisticians made it a point to not only analyze the overall effects of the model, but also analyze the appropriateness of the conclusions drawn from the model assumptions and structures. After much review, these researchers ended up concluding that regardless of the size of the dataset, causal inferences are not always appropriate conclusions of a complex analysis of observational data, even given propensity adjustment. This is because there is almost no way to consider and control for all possible confounding and contributing factors to a treatment outcome of an observational study. This is an important aspect of observational study structure that needs to be constantly considered and controlled whenever an analysis is conducted. On the other hand, another interesting outcome of this study was that the propensity scores created were effectively utilized in the analysis of this type of dataset; therefore, there exists little reason that this type of analysis shouldn't be able to be utilized in a complex sample of similar structure.

Compounding on this conclusion, one must also consider that there still exists some methodological challenges of propensity score utilization in some sampling methods (such as complex survey sampling) that could have severe impacts on result interpretations if not appropriately identified, approached, and controlled (Pan & Bai, 2015). A big take-away point for this challenge is that before consideration of propensity score utilization for model adjustment, the researcher must consider the assumptions and theoretical implications of their sample to make sure that propensity score theory matches the research question and would be a viable option for bias control. In the case of this study, appropriate measures were taken and extra care given to the choice of variables in order to ensure methodological compatibility with propensity score utilization.

After the decision to utilize propensity scores has been made, the steps needed for score

creation and method utilization must then be implemented. Matching, stratification, and regression adjustment are all statistical techniques commonly employed after propensity score creation. Alone, the interpretive possibilities of these techniques can be severely limited given small covariate inclusions; however, when propensity score techniques are employed in conjunction with one of these statistical processes, the covariate information needed is summarized into a single score, thus diminishing this limitation and opening up the results to interpretation. It is, therefore, beneficial to utilize propensity scores in addition to one of these statistical techniques. For the purposes of this study, two types of regression adjustment were considered along with stratification and matching techniques. Simple regression adjustment in the form of a covariate, weighted regression adjustment through use of the inverse probability of treatment weights, and quintile-stratified model adjustment were the chosen propensity score utilization strategies employed in this study. Matching was excluded based on the methodological concerns over the utilization of an ordinal predictor variable and limited sample size.

METHODS

DATA

The data for this study is meant to represent 5 years of medical student data. The data was created to imitate actual data collected school on American medical students for use in a theoretical exploration. In order to protect the identity and implications of this analysis, this study does not include any actual data. Any results presented in this paper should only be taken as a theoretical exploration and not a final conclusion of the model.

STATISTICS

A predefined group of variables from dataset in question were used in the descriptive explanation. A smaller subset of these variables were used in the final analysis. Demographic and pre-medical variables include: age at admissions, class year, gender, military branch, college science GPA, college total GPA, MCAT Biologic Science score, MCAT Physical Science score, MCAT Verbal Reasoning Score, and MCAT Total Score. Preclerkship (first 2 years of medical school) variables include: CPR NBME final score, Endocrinology & Reproduction NBME final score, Fundamentals NBME final score, Gastro-intestinal NBME final score, Musculoskeletal NBME final score, and Neurology NBME final score. Clerkship (3rd year of medical school) variables include: Family Medicine NBME score, Internal Medicine NBME score, Pediatrics NBME score, Psychiatry NBME score, Surgery NBME score, Ambulatory Round, Ward Round, General Surgery Round, Special Surgery Round, Family Medicine Round, Internal Medicine Round, Obstetrics and Gynecology Round, Pediatrics Round, Psychiatry Round, Internal Medicine and Psychiatry Block (and Sequence), Surgery and Obstetrics Block (and Sequence), and Family Medicine and Pediatrics Block (and Sequence). Outcome variables of interest include the clerkship NBMEs and USMLE Step 1 and Step 2 CK exam scores.

All NBME and USMLE examinations are standardized examinations. Demographic, GPA, and MCAT data were all gathered at admissions. GPA and MCAT data were averaged in the event of multiple entries (ie. Taking MCAT more than once or having multiple degrees).

For descriptive purposes, we chose to bin the USMLE variables and run the frequency and means procedures on these binned variables. The HPBin procedure was used in order to do this:

```
/*Bucket Binning */
proc hpbin data=bindataset numbin=4 bucket computestats
output=bucketbin;
```

```
      input USMLE_Step1_1st_Time_Score; input
USMLE_Step2_1st_Time_Score;
      ods output mapping=result_bucket;
      code file='C:\Users\dschreiber-gregory\Desktop\Complete Projects
[NOT PUBLIC]\[2020.03] SAS Global Forum - Propensity Score
Paper/Bucket_BinCode.sas';
run;

/* Winsorized Binning */
proc hpbin data=bindataset numbin=4 winsor computestats
output=winsorbin;
      input USMLE_Step1_1st_Time_Score; input
USMLE_Step2_1st_Time_Score;
      ods output mapping=result_winsorized;
      code file='C:\Users\dschreiber-gregory\Desktop\Complete Projects
[NOT PUBLIC]\[2020.03] SAS Global Forum - Propensity Score
Paper/Winsor_BinCode.sas';
run;

/*Pseudo-Quantile Binning*/
proc hpbin data=bindataset numbin=4 pseudo_quantile computehist
computequantile output=pseudobin;
      input USMLE_Step1_1st_Time_Score; input
USMLE_Step2_1st_Time_Score;
      ods output mapping=result_pseudo;
      ods output histogram=histo_pseudo;
      code file='C:\Users\dschreiber-gregory\Desktop\Complete Projects
[NOT PUBLIC]\[2020.03] SAS Global Forum - Propensity Score
Paper/Pseudo_BinCode.sas';
run;

/*Bin Comparison*/
proc hpbin data=bindataset woe bins_meta=result_bucket;
      target Class/level=nominal;
run;

proc hpbin data=bindataset woe bins_meta=result_winsorized;
      target Class/level=nominal;
run;

proc hpbin data=bindataset woe bins_meta=result_pseudo;
      target Class/level=nominal;
run;

data &dataset;
      set &dataset;
      %include 'C:\Users\dschreiber-gregory\Desktop\Complete Projects
[NOT PUBLIC]\[2020.03] SAS Global Forum - Propensity Score
Paper/Bucket_BinCode.sas';
run;
```

Descriptive statistics were run using Proc Freq and Proc Means procedures. These statistics were used in order to get a handle on the distribution of the data and to help eliminate variables that would not contribute meaningfully to the final model.

```
      /* Descriptives for USMLE */

      proc freq data=&dataset;
```

```
      tables (IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block) *
BIN_USMLE_Step1_1st_Time_Score;
      tables (IntMed_Psych_Sequence Surgery_OB_Sequence
FamMed_Ped_Sequence) * BIN_USMLE_Step1_1st_Time_Score;
run;

proc freq data=&dataset;
      tables (Class Gender Service) * BIN_USMLE_Step1_1st_Time_Score /
chisq;
run;

proc sort data=&dataset;
      by BIN_USMLE_Step1_1st_Time_Score;
run;

proc means data=&dataset;
      by BIN_USMLE_Step1_1st_Time_Score;
      var Age_Admissions CollegeBCPMGPA1 CollegeTOTALGPA
MCATBiologicScience1 MCATPhysicalScience1 MCATVerbalReasoning1
MCATTOTAL1
            CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final;
run;

/* Descriptives for Block */

proc freq data=&dataset;
      tables (Class Gender Service) * IntMed_Psych_Block / chisq;
      tables (Class Gender Service) * Surgery_OB_Block / chisq;
      tables (Class Gender Service) * FamMed_Ped_Block / chisq;
run;

proc sort data=&dataset;
      by IntMed_Psych_Block;
run;

proc means data=&dataset;
      by IntMed_Psych_Block;
      var Age_Admissions CollegeBCPMGPA1 CollegeTOTALGPA
MCATBiologicScience1 MCATPhysicalScience1 MCATVerbalReasoning1
MCATTOTAL1
            CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final;
run;

proc sort data=&dataset;
      by Surgery_OB_Block;
run;

proc means data=&dataset;
      by Surgery_OB_Block;
      var Age_Admissions CollegeBCPMGPA1 CollegeTOTALGPA
MCATBiologicScience1 MCATPhysicalScience1 MCATVerbalReasoning1
MCATTOTAL1
            CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final;
run;
```

```
proc sort data=&dataset;
      by FamMed_Ped_Block;
run;

proc means data=&dataset;
      by FamMed_Ped_Block;
      var Age_Admissions CollegeBCPMGPA1 CollegeTOTALGPA
MCATBiologicScience1 MCATPhysicalScience1 MCATVerbalReasoning1
MCATTOTAL1
            CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final;
run;
```

RESULTS: UNIVARIATE AND EXPLORATIVE MULTIVARIATE ANALYSES

A multivariate logistic regression analysis for an adjusted model was then conducted to test whether or not and to what extent scores on the individual shelf examinations while controlling for clerkship sequence and covariates helped explain the variation in USMLE exam score. A multivariate logistic regression analysis was then conducted to test whether or not and to what extent clerkship sequence helped explain the variation in individual shelf exam scores.

```
proc glmselect data=&dataset;
      title1 'Stepwise  Model Fit Selection';
      class IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block Class
Gender Service;
      model USMLE_Step1_1st_Time_Score = Age_Admissions CollegeBCPMGPA1
CollegeTOTALGPA MCATBiologicScience1 MCATPhysicalScience1
MCATVerbalReasoning1 MCATTOTAL1
                  CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final
                  IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block
Class Gender Service
                  FamMed_NBME IntMed_NBME ObGyn_NBME Ped_NBME
Psych_NBME Surgery_NBME/ selection=stepwise;
run;

proc logistic data=&dataset;
      class IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block Class
Gender Service;
      model USMLE_Step1_1st_Time_Score = Age_Admissions CollegeBCPMGPA1
MCATBiologicScience1 MCATPhysicalScience1
                  CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final
                  IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block
Class Gender Service
                  FamMed_NBME IntMed_NBME ObGyn_NBME Ped_NBME
Psych_NBME Surgery_NBME/ rsq;
run;
```

Through this analysis, subsequent propensity scores were also produced and output into a variable with intent to be used as either a covariate adjustment or weight in the final model. Additional histogram plots were also produced and assessed. Through evaluation of these histogram plots, it is clear that there exists significant overlap in the covariate distributions between each of the groups, indicating that the groups are now comparable and ready for appropriate inclusion into the model. For the record, adjusted odds ratio scores were also produced reviewed.

RESULTS: PROPENSITY SCORE ADJUSTMENTS

METHOD BEHIND PROPENSITY SCORE CREATION

As with other statistical procedures, the validity of a propensity score must adhere to a set of assumptions. When using propensity scores for causal inference (or the reduction of selection bias in an observational study), the following assumptions must be met:

- Stable Unit Treatment Value Assumption (SUTVA): this assumption states that the potential outcomes for any subject must not vary from the intervention assigned to another subject. For each subject considered, each intervention level must be the same as all other subjects and must lead to the same potential outcomes.
- Positivity: this assumption states that the probability of an assignment to an intervention for each subject must strictly exist between 0 and 1.
- Unconfoundedness: this assumption states that the assignment to a treatment for each subject must be independent of the potential outcomes, given a set of covariates from before intervention

If these assumptions are met, then a propensity score may be used as a balancing score, meaning that the treatment assignment is independent of the potential outcome, given the propensity score.

The logistic model provides a description of the relationship of several independent variables to a dichotomous dependent variable. Furthermore, logistic regression is used to predict the probability of an event occurring as a function of a set of independent variables (continuous and/or dichotomous). The logistic model can be represented as such:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i x_i)}}$$

Propensity scores can be easily created through the LOGISTIC or PSMATCH procedure in SAS®. In the case used and steps described in this paper, the dependent variable is treatment group (USMLE Score) and the independent variables are clerkship blocks and other demographic and educational factors. The GENMOD procedure for generalized linear models may also create propensity scores by using the OUTPUT statement and keyword PREDICTED.

The PSMATCH procedure was introduced in SAS/STAT v14.2. It was included in order to provide a variety of tools for specific practice of performing propensity score analysis. The PSMATCH procedure is designed to reduce the effects of confounding in nonrandomized trials or observational studies where the subjects are not randomly assigned to the treatment and control groups. The following methods for using the propensity scores to adjust the data for valid estimation of treatment effect are available through this procedure:

- Inverse probability of treatment weighting and weighting by the odds.
- Stratification of observations that have similar propensity scores. In a subsequent outcome analysis, the treatment effect can be estimated within each stratum, and the estimates can be combined across strata to compute an average treatment effect.
- Matching treated unit with one or more control units that have a similar value of the propensity score. Methods of matching include:
    - fixed ratio matching
    - variable ratio matching
    - full matching

PSMATCH can also provide various plots for balance assessment. Included plots are:

- cloud plots, which are scatter plots in which the points are jittered to prevent overplotting
- box plots for continuous variables
- bar charts for classification variables
- a standardized differences plot that summarizes differences between the treated and control groups

The PSMATCH procedure then saves propensity scores and weights in an output data set that contains a sample that has been adjusted either by weighting, stratification, or matching (whichever is chosen by the user). Additionally, if the sample is stratified, you can save the strata identification in the output data set. If the sample is matched, you can save the matching identification in the output data set.

SIMPLE REGRESSION ADJUSTMENT

In the first approach to utilizing a propensity score adjustment, the produced propensity score was added as is into the model as a covariate for a simple regression adjustment.

```
/* Regression Adjustment - Simple */
proc logistic data = AllPropen;
      class IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block Class
Gender Service;
      model USMLE_Step1_1st_Time_Score = Age_Admissions CollegeBCPMGPA1
MCATBiologicScience1 MCATPhysicalScience1
                  CPR_NBME_Final EndoRepro_NBME_Final Fund_NBME_Final
GI_NBME_Final MSK_NBME_Final Neuro_NBME_Final
                  IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block
Class Gender Service
                  FamMed_NBME IntMed_NBME ObGyn_NBME Ped_NBME
Psych_NBME Surgery_NBME / lackfit rsq;
      title 'Propensity Scores Adjusted';
run;
```

INVERSE PROBABILITY OF TREATMENT WEIGHT ADJUSTMENT

In the second approach to utilizing a propensity score adjustment, the produced propensity score was then recalculated as the inverse of the propensity weight and added back into the model as a weight adjustment, otherwise referred to as the inverse probability of treatment weight (Hogan and Lancaster, 2004). Since this calculation assumes only two levels of a predictor variable (the propensity of which was calculated), then the Surgery NBME variable needed to be recoded as dichotomous for the weight calculation. The recode thus identified Surgery NBME score as “Low/Average” (the medical student scoring lower than the 65th percentile) or “High” (the medical student scoring in at least the 65th percentile). Inverse probability of treatment weight was calculated based on this association and added into the model as a weight.

The treatment selection model above modeled the propensity to score higher or lower on the Surgery NBME exam. For those medical students who scored higher on the Surgery NBME exam, the propensity score would be 1-ps and the propensity score weight would be the inverse of 1-ps.

```
/* Regression Adjustment - Inverse Probability of Weights */
```

```
data AllPropen;
      set AllPropen;
      If Surgery_NBME_Grade=0 then ps_weight=(1/prob);
            else ps_weight=(1/(1-prob));
run;

ods graphics on;
proc psmatch data=&dataset region=allobs;
   class Surgery_NBME_Grade IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block;
   psmodel Surgery_NBME_Grade (Treated='1')= IntMed_Psych_Block
Surgery_OB_Block FamMed_Ped_Block;
   psweight weight=atewgt nlargestwgt=6;
   assess lps var=(IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block)
          / varinfo plots=(barchart boxplot(display=(lps BMI))
wgtcloud);
   id Surgery_OB_Block;
   output out(obs=all)=OutEx1 weight=_ATEWgt_;
run;
```

STRATIFICATION

Stratification, subclassification or binning using propensity scores involves grouping subjects into classes or strata based on the subject's observed characteristics. Once the propensity scores are calculated, subjects are placed into strata (Cochran states that 5 strata can remove 90% of the bias) with the idea that subjects in the same stratum are similar in the characteristics used in the propensity score development process. The tutorial by D'Agostino details how to perform this technique. Briefly, quintiles are used to group subjects into five strata after making sure that there is adequate propensity scores overlap between the treatment groups. To prove that the propensity scores removed any bias due to differences in covariates between treatment groups, t-tests or chi-square tests are conducted before and after propensity score creation. Finally, outcomes and treatment effects can be assessed using models while adjusting for the propensity scores. Continuing with the example and code above, subjects are divided into 5 classes based on the common propensity score overlap using the RANK procedure. Checking for difference between treatment group before and after stratifying subjects by propensity scores can be done using PROC FREQ, PROC TTEST and PROC GLM.

```
/* Stratification */
proc rank data=AllPropen groups=5 out=r;
      var prob;
      ranks rnks;
run;

data quintile;
      set r;
      quintile=rnks+1;
run;

proc contents data=quintile;
run;

proc freq data=quintile;     /* Check for differences in groups before
propensity score */
```

```
      tables Surgery_NBME_Grade*(IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block) / chisq;
run;

proc logistic data=quintile; /* Check for differences in groups while
adjusting for propensity scores */
      class Surgery_NBME_Grade IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block / param=ref;
      model Surgery_NBME_Grade = IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block quintile / lackfit rsq;
                  oddsratio Surgery_NBME_Grade / cl=wald;
run;
quit;

proc logistic data=quintile; /* Check for differences in groups while
adjusting for propensity scores */
class Surgery_NBME_Grade IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block / param=ref;
model Surgery_NBME_Grade = IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block quintile / lackfit rsq;
                  oddsratio Surgery_NBME_Grade / cl=wald;
                  oddsratio IntMed_Psych_Block / cl=wald;
                  oddsratio Surgery_OB_Block / cl=wald;
                  oddsratio FamMed_Ped_Block / cl=wald;
                  oddsratio quintile / cl=wald;
run;
quit;

ods graphics on;
proc psmatch data=&dataset region=allobs;
   class Surgery_NBME_Grade IntMed_Psych_Block Surgery_OB_Block
FamMed_Ped_Block;
   psmodel Surgery_NBME_Grade (Treated='1')= IntMed_Psych_Block
Surgery_OB_Block FamMed_Ped_Block;
   strata nstrata=5 key=treated stratumwgt=total;
   assess ps var=(IntMed_Psych_Block Surgery_OB_Block FamMed_Ped_Block)
          / varinfo plots=(barchart cdfplot);
   output out(obs=all)=OutEx2;
run;
```

The results of this particular analysis were similar to the above proposed tables as they were able to show minimal differences between groups when subclassifying subjects. Outcomes would then be able to be compared within the five subclasses or averaged to report for the overall treatment groups.

MATCHING

Another common method to balance on covariates is matching groups by propensity scores. With this method, once the propensity score is calculated, participants are then matched on this single score instead of the traditional direct matching technique by one or more covariates. The main disadvantage of this method is that the resulting matches could be incomplete or inexact. In other words, subjects may end up being excluded from the final analysis due to difficulty in finding a match. Fortunately, there is a way to reduce this bias. The process of reducing the bias of matching propensity scores is thoroughly explained in a series of papers authored by Lori Parsons, of which is referenced at the end of this paper. Her papers include an explanation of each proposed procedure and subsequent macro code for performing case-control matches using a greedy matching algorithm. Matching was excluded

from this review based on the methodological concerns over such a small sample size of medical students (post curriculum reform) and the use of an ordinal predictor variable. However, it is worthy to note the possibility of its usage in a similar, more complete study

MODEL COMPARISON

In order to compare the effectiveness and fit of the multivariate logistic regression model before propensity adjustment to the multivariate logistic regression models after propensity adjustment, model fit statistics and r-square values produced by each model were reviewed. In review of these statistics, it is worthy to note that the Cox-Snell r-square and max-rescaled r-squares are default predictive power calculations in SAS. The max-rescaled r-square is the one recommended for use in predictive power comparisons between the models. This adjusted version is a recalculated r-square produced by SAS as a solution to the upper-level boundary issues identified in the original Cox-Snell r-square calculations. For the purpose of this study, the max-rescaled r-square statistics will be the ones reviewed.

In addition to this, goodness of fit tests are also produced for model comparisons of fit. The Akaike Information Criterion (AIC), Schwarz Criterion (SC), and -2 log likelihood are default goodness of fit productions for the SAS logistic procedure. The -2 log likelihood statistic has a chi-square distribution under the null hypothesis (in other words, it tests whether all explanatory variables in the model have zero significance) and produces a p-value for statistical comparison. The AIC and SC statistics are two adjustments for -2 log likelihood statistic based on the number of terms in the model and the number of observations that are being used. The AIC and SC statistics are of primary interest and will be used in the comparison of the different models. As a rule, lower values of the AIC and SC statistics indicate a more appropriate model. In addition to these default statistics, the Hosmer-Lemeshow test was also conducted in order to check for the overall fit of each model and to serve as a guide for structuring future iterations of this study. The Hosmer-Lemeshow goodness-of-fit test is specifically designed for binary response models, such as the one in this study. Through employment of this test the participants are divided into approximately ten groups of about the same size based on the percentiles of estimated probabilities. The discrepancies between the observed and expected number of observations within these groups are then summarized through use of the Pearson chi-square statistic, then compared to a chi-square distribution with t degrees of freedom (t = number of groups – n, with n=2 as default).

In reviewing the test results, a small “significant” p-value suggests that the fitted model is not acceptable. A new model with additional covariates or different predictor variables would then need to be explored. However, Allison states in his 2014 lecture at SAS Global Forum that the Hosmer-Lemeshow test, though a decent measure of fit, is not a perfect measure. It does have some serious problems that need to be addressed before it can become a gold standard. One such problem is that the results that it produces are highly dependent on the number of groups specified for the model (as stated earlier, this number is ten by default in SAS). This would not be much of a problem if there existed some theory to guide in the appropriate calculation of these groups, however, no such theory exists, leaving the decision of group number either up to the statistician or the default values of the program. Another note to consider when reviewing Hosmer-Lemeshow test results is that the test itself was developed for use in small datasets; therefore, when applying this test to a larger sample size, the overall interpretive ability of the test is compromised (Kramer & Zimmerman, 2007). Keeping these points in mind, the Hosmer-Lemeshow test results will be reviewed for this study; however, their implications will not impact the validity of the study, rather, decisions for possible additional covariate exploration and sample size adjustments in future studies will be explored in response to undesirable Hosmer-Lemeshow test results.

CONCLUSION

In conclusion, propensity score utilization is not only possible for an educational study such as this, but actually added to the accuracy and fit of the model. Propensity score utilization for project such as this, however, needs to be carefully considered and utilized if chosen. Propensity score assumptions, model assumptions, missing information, appropriateness of variable type and structure, and many more considerations need to be reviewed before and during implementation.

REFERENCES

Allison, P. D. (2014). Measures of Fit for Logistic Regression. Proceedings for SAS Global Forum 2014, Washington, DC, 1485-2014.

Committee for Proprietary Medicinal Products (CPMP). (2003). Points to Consider on Adjustment for Baseline Covariates. The European Agency for the Evaluation of Medicinal Products. Retrieved from http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/06/WC500144946.pdf.

D'Agostino, R.B., Sr., & Kwan, H. (1995). Measuring Effectiveness: What to Expect Without a Randomized Control Group. Medical Care. 195(33): AS95-AS105.

D'Agostino R.B., Jr, & D'Agostino R.B., Sr. (2007). Estimating Treatment Effects Using Observational Data. JAMA, 297(3): 314-316.

D'Agostino, R.B. (1998). Tutorial on Biostatistics: Propensity Score Methods for Bias Reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine, 17: 2265-2281.

Faries, D., Zhang, X., Kadziola, Z., Siebert, U., Kuehne, F., Obenchain, R., Haro, J. M. (2019) Real World Health Care Data Analysis – Causal Methods and Implementation Using SAS. SAS Press: Cary, NC.

Hogan, J.W., & Lancaster, T. (2004). Instrumental variable and propensity weighting for causal inference from longitudinal observational studies. Statistical Methods in Medical Research, 13: 17-48.

Kramer, A. A., & Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Critical Care Medicine. 35(9): 2052-6.

Lanehart, R., Rodriguez de Gil, P., Kim, E. S., Bellara, A. P., Kromrey, J. D., & Lee, R. S. (2012). Propensity Score Analysis and Assessment of Propensity Score Approaches Using SAS® Procedures. Proceedings of SAS Global Forum 2012. Orlando, FL, 314-2012.

Obenchain, R.L., & Melfi, C.A. (1997). Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies. Proceedings of the American Statistical Association: Biopharmaceutical Section, Anaheim, CA, 297-306.

Pan, W., & Bai, H. (2015). Propensity score analysis. New York: The Guilford Press.

Pasta, D. J. (2000). Using Propensity Scores to Adjust for Group Differences: Examples Comparing Alternative Surgical Methods. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference 2000, Indianapolis, IN, 261-25.

Parsons, L. (2000). Using SAS® Software to Perform a Case Control Match on Propensity Score in an Observational Study. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference 2000, Indianapolis, IN, 214-26.

Rosenbaum P.R. and Rubin D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41-55.

Ross, M. E., Kreider, A. R., Huang, Y., Matone, M., Rubin, D. M., & Localio A. R. (2015). Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures. American Journal of Epidemiology, 181(12): 989-95.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Statistical Medicine. 26(1): 20-36.

SAS Institute Inc. 2013. "SAS/STAT User's Guide". SAS OnlineDoc® 9,4. Cary, NC: SAS Institute Inc. https://support.sas.com/documentation/cdl/en/procstat/65544/PDF/default/procstat.pdf.

Stuart, E. A., DuGoff, E., Abrams, M., & Salkever, D. (2013). Estimating Causal Effects in Observational Studies Using Electronic Health Data: Challenges and (some) Solutions. EGEMs (Generating Evidence & Methods to improve patient outcomes). 1(3): 4.

CONTACT INFORMATION

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

Deanna N Schreiber-Gregory, MS
d.n.schreibergregory@gmail.com