Paper 689-2017

A Practical Guide to Getting Started with Propensity Scores

Thomas Gant, Keith Crowland
Data & Information Management Enhancement (DIME)
Kaiser Permanente

ABSTRACT

This paper gives tools to begin using propensity scoring in SAS® to answer research questions involving observational data. It is for both those attendees who have never used propensity scores and those who have a basic understanding of propensity scores, but are unsure how to begin using them in SAS. It provides a brief introduction to the concept of propensity scores, and then turns its attention to giving you tips and resources that will help you get started. The paper walks you through how the code in the book "Analysis of Observational Health Care Data Using SAS®", which is published by the SAS Institute, is used to analyze how a health care treatment impacted an associated health care outcome. It details how propensity scores are created and how propensity score matching is used to balance covariates between treated and untreated observations. With this case study in hand, you will feel confident that you have the tools necessary to begin answering some of your own research questions using propensity scores.

A BRIEF INTRODUCTION TO PROPENSITY SCORES

A propensity score is the conditional probability that a subject receives “treatment” given the subject’s observed covariates. The goal of propensity scoring is to mimic what happens in randomized controlled trials (RCT’s) by balancing observed covariates between subjects in control and treatment study groups (Faries, Leon, Haro, Obenchain, 2010).

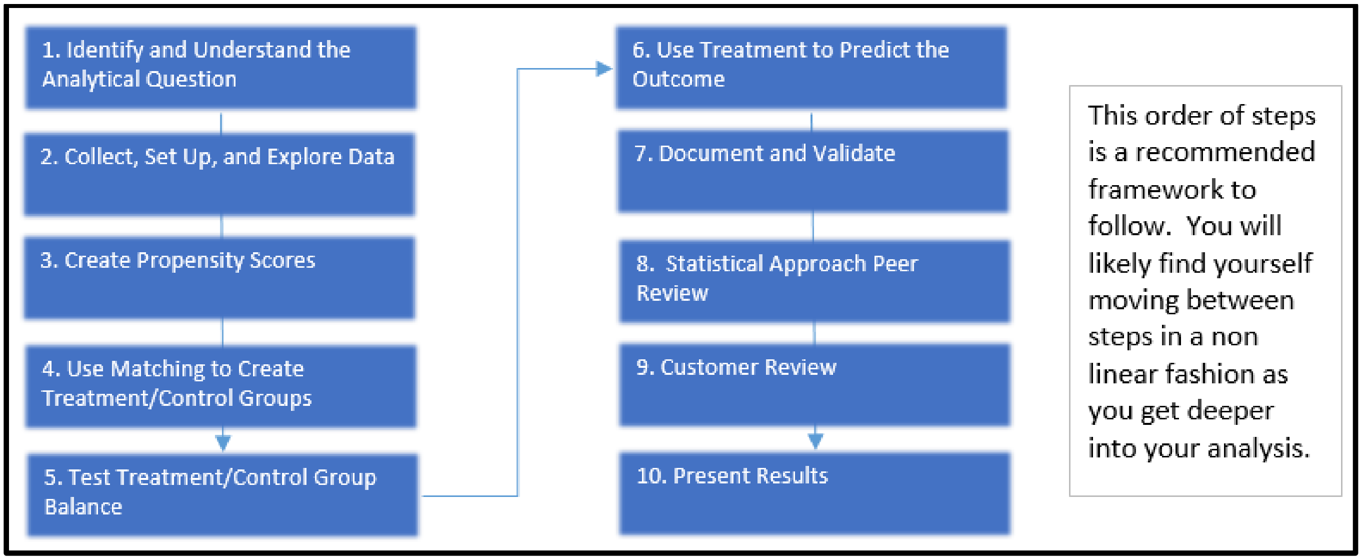
Randomized controlled trials are considered the gold standard when evaluating a treatment’s effectiveness. In a randomized controlled trial, subjects are randomly placed into a treatment and a control group. The treatment group receives the treatment and the treatment outcomes are evaluated versus the control group outcomes. Due to randomization, an RCT has no selection bias when splitting subjects up into the control and treatment groups. Randomization balances both observed and unobserved characteristics between the two subject groups. Thus it accounts for all possible “confounding variables”. Confounding variables are independent variables other than the treatment variable that are correlated to the outcome of the study. Unaccounted for confounding variables prevent us from measuring the true impact a treatment has on an outcome.

Randomized controlled trials can be expensive, resource intensive to perform, and in certain circumstances unethical to perform. For these reasons, data scientists often rely on observational data. The challenge with observational data is that treatments are not applied randomly, leading to selection bias and confounding variables. For example, if you study the impact of a heart disease medication, and you compare the outcome of all those who receive the medication versus all of those who didn’t, you would likely have a huge level of selection bias, as those who receive the medication are likely to have higher blood pressure, a higher BMI, diabetes, etc. Improved confounding variable balance between treatment and control groups can be achieved by matching observations from each group based on the propensity score, which in this case would be the probability that a patient received the medicine given the observed covariates. Propensity score analysis seeks to isolate the treatment as the only difference between our treatment and control groups.

STEPS TO PERFORM A PROPENSITY SCORE ANALSYSIS

The purpose of this paper is to give you the tools you need to begin performing propensity score analyses. It is geared towards hands-on learning. It will take you through each step you need to carry out a successful analysis using SAS®, while highlighting common pitfalls to avoid. As mentioned above, it will borrow heavily from the concepts and code in the first three chapters of “Analysis of Observation Health Care Data Using SAS®", published by the SAS institute.

To help accomplish this, the paper will walk you through a case study performed at Kaiser Permanente regarding patients who received a treatment and to what degree, if at all, this treatment led to a more adverse outcome. In this case study, the analysis moves through a series of steps which are important to completing a sound analysis. These steps are shown in Figure 1.


1. Identify and Understand the Analytical Question
2. Collect, Set Up, and Explore Data
3. Create Propensity Scores
4. Use Matching to Create Treatment/Control Groups
5. Test Treatment/Control Group Balance
6. Use Treatment to Predict the Outcome
7. Document and Validate
8. Statistical Approach Peer Review
9. Customer Review
10. Present Results

This order of steps is a recommended framework to follow. You will likely find yourself moving between steps in a non linear fashion as you get deeper into your analysis.

Figure 1. The recommended steps to take when completing a propensity score analysis

Each of the numbered steps in Figure 1 correspond to the titled sections in the paper that follow. The paper covers all steps, with more emphasis placed on steps 3 through 5, which involve creating propensity scores, creating treatment and control groups with one to one propensity score matching, and testing for balance between the treatment and control groups.

As you go through model validation, statistical approach peer review, and customer review, adjustments are made to the analysis which require a fresh look at your approach to the question at hand. This review process that occurs in steps 7-9 is vital to have confidence in the final analysis, and multiple runs through the steps above are part of the analytical process. Thorough research of the business problem, along with more experience using propensity scores, should reduce the number of times you need to go through the steps above before completing your analysis.

Note that this paper uses the term “customer” frequently. In the context of this paper, a “customer” is the individual requesting the analysis, whether that individual sits within or outside of your organization.

STEP 1: IDENTIFY AND UNDERSTAND THE BUSINESS QUESTION

Due to the proprietary nature of this analysis, we have generalized our case study and do not detail the specific treatment or outcome measured. This is appropriate as our primary purpose in this paper is to provide a practical guide to performing propensity score analysis, and the details of the case study are not required to accomplish this.

Our customer has reason to believe that a treatment leads to an adverse health outcome. To determine whether the treatment has a negative impact, the customer requested several years’ worth of inpatient encounter data, along with information on whether the patient received the treatment or not. A higher value in our outcome metric indicates an adverse result. An examination of the raw data allowed the customer to determine that the patients who receive treatment have a higher outcome value than those that do not.

The customer is savvy enough to have recognized that the problem may be more complex than calculating an average of the outcome for the treated and untreated groups, and requested that we validate the findings. We recognize that there are many variables that can affect the outcome and that we need to account for as many of these covariates as possible.

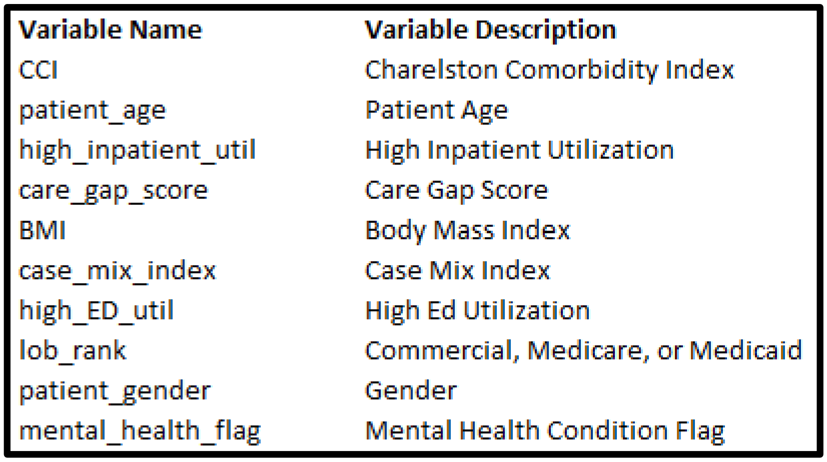
This is where the importance of interviewing customers and other subject matter experts comes into play. Understanding as much of the problem up front is hugely beneficial, and it will limit the number of times that each of the steps highlighted in Figure 1 above need to be completed. An omitted covariate or an uninformed decision regarding outliers can change the results of the study. Getting the viewpoint of those with the most knowledge of patients who receive the treatment allows us to make our most informed attempt to include the most important covariates in our study.

Our interviews and our existing knowledge helped us to determine that a propensity score based analysis is appropriate to answer our analytical question. The reasons we made this determination are:

- This study is based on observational data and thus has selection bias; our treatment group is not assigned the treatment randomly.
- There are numerous confounding variables; health metrics like patient age and patient body mass index (BMI), among other variables, have the potential to impact our outcome and be confounding variables.

STEP 2: COLLECT, IDENTIFY, AND EXPLORE DATA

Through our interview process in step 1, several potential confounding variables that could affect our outcome are identified to be included in our analysis. These variables are shown in Table 1.

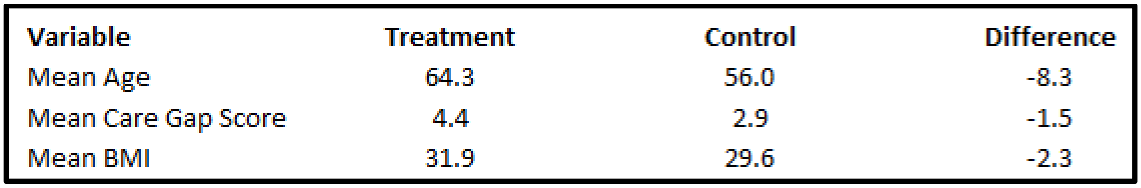


Variable Name	Variable Description
CCI	Charelston Comorbidity Index
patient_age	Patient Age
high_inpatient_util	High Inpatient Utilization
care_gap_score	Care Gap Score
BMI	Body Mass Index
case_mix_index	Case Mix Index
high_ED_util	High Ed Utilization
lob_rank	Commercial, Medicare, or Medicaid
patient_gender	Gender
mental_health_flag	Mental Health Condition Flag

Table 1. Variable name and variable description of potential confounding variables in our study

We organized our data into a single dataset with all the potential confounding variables that we identified, along with a column that indicates whether the patient received the treatment.

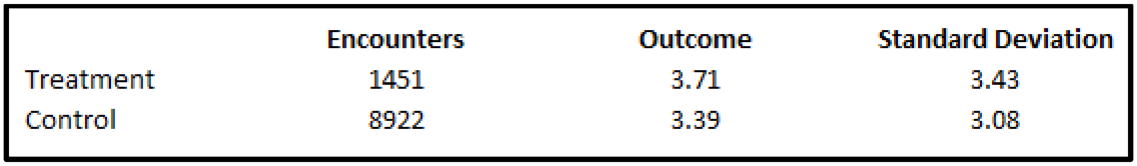
We need to account for any of the predictors that are correlated with the outcome. For example, if patients who receive the treatment are more likely to be older, how would we know if their negative outcomes are because of their treatment or because of their older age? In Table 2, we see that patients who received the treatment do have a higher BMI, are older, and have more care gaps.



Variable	Treatment	Control	Difference
Mean Age	64.3	56.0	-8.3
Mean Care Gap Score	4.4	2.9	-1.5
Mean BMI	31.9	29.6	-2.3

Table 2. The mean age, care gap score, and BMI for patients in out treatment and control groups

We also validated the customer’s initial findings that patients receiving the treatment have a higher outcome value (and thus a more negative health impact) than patients not receiving treatment. Table 3 shows that the mean outcome for patients receiving the treatment is 3.71, while the mean outcome for patients not receiving the treatment is 3.39.



	Encounters	Outcome	Standard Deviation
Treatment	1451	3.71	3.43
Control	8922	3.39	3.08

Table 3. Encounter count and health outcome value for our treatment and control groups

In our study, to test for a significant difference between the mean outcome values for our treatment and control groups, a negative binomial model was used with the treatment (yes or no) as the independent variable and the outcome as the dependent variable. Patients receiving treatment had a statistically significant higher outcome value than those not receiving treatment. Remember in this study a higher value in our outcome metric reflects a more adverse health outcome. Unfortunately, because this is an observational study, we do not know if the higher outcome value is because the patient received the treatment or because confounding variables have not been considered in the analysis.

Note that when testing for significance between two means you will want to apply a model that fits the distribution of your data. Our outcome values have a negative binomial distribution, which is why we chose that model. The code used to create a negative binomial model is shown later in this paper.

STEP 3: CREATE PROPENSITY SCORES

Remember that the goal of propensity scores is to balance observed covariates between subjects from the treatment and control groups to imitate what happens in a randomized study (Faries, Leon, Haro, Obenchain, 2010). In our case, that means we would want our treatment and control groups to have a similar mean age, mean BMI, etc., just as if we are doing a randomized controlled trial.

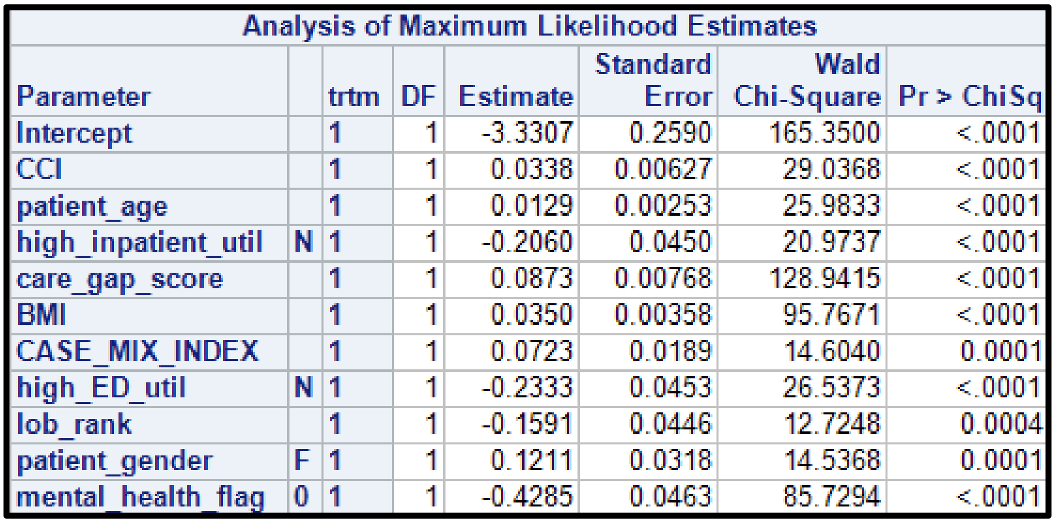
Logistic Regression is often used to develop propensity scores, which in our case study represent the probability that a patient receives the health treatment based on the subject's observed covariates. Thus, we created a logistic model in which the variables we would like to balance between, our treatment and control groups, are used as our predictors of treatment, our dependent variable in the logistic regression model.

We are not overly interested in how well our logistic regression model predicts whether a patient receives treatment or not. Typically, in logistic regression, we would want to maximize our 'C Score' in the SAS output, which is a measure of discriminatory predictive power (Westreich, Cole, Funk, Brookhart, Sturmer, 2012). In a propensity score analysis, it is more important that we include all predictor variables in the logistic regression model that are correlated with our health outcome. This means that we may include variables that lower our logistic model's predictive power of the treatment. If these variables are related to the health outcome, then they should be balanced between our treatment and control groups, and used to create our propensity scores.

We modified the code on page 64 in "Analysis of Observational Health Care Data Using SAS" to fit our data and create propensity scores:

```
/* Create Propensity Scores */
proc logistic data=inpatient_encounters;
   class high_inpatient_util high_ED_util lob_rank patient_gender
   mental_health_flag
   model treatment = cci patient_age high_inpatient_util
   care_gap_score BMI case_mix_index high_ED_util lob_rank
   patient_gender mental_health_flag/
   link=glogit rsquare;
   output out = ps_los pred = ps xbeta=logit_ps;
   /* Output the propensity score and the logit of the propensity
score */
run;
```

The proc logistic code above will add two fields to our dataset, propensity score and the logit of the propensity score. We will use both fields when we perform propensity score matching in the next step. Table 4 below is partial output from the code, and gives all variables that we included in our logistic model and that we would like to balance between the control and treatment groups. You can see that in this case each variable was significant in predicting whether treatment is received.



Analysis of Maximum Likelihood Estimates

Parameter		trtm	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1	-3.3307	0.2590	165.3500	<.0001
CCI		1	1	0.0338	0.00627	29.0368	<.0001
patient_age		1	1	0.0129	0.00253	25.9833	<.0001
high_inpatient_util	N	1	1	-0.2060	0.0450	20.9737	<.0001
care_gap_score		1	1	0.0873	0.00768	128.9415	<.0001
BMI		1	1	0.0350	0.00358	95.7671	<.0001
CASE_MIX_INDEX		1	1	0.0723	0.0189	14.6040	0.0001
high_ED_util	N	1	1	-0.2333	0.0453	26.5373	<.0001
lob_rank		1	1	-0.1591	0.0446	12.7248	0.0004
patient_gender	F	1	1	0.1211	0.0318	14.5368	0.0001
mental_health_flag	0	1	1	-0.4285	0.0463	85.7294	<.0001

Table 4. The results of our logistic models which is used to create our propensity scores

The variables we would like to balance between the treatment and the control group don't necessarily have to be significant in predicting treatment. As mentioned above, to reduce the chance of hidden bias, it is important to be more inclusive when determining which variables should be in the model. Hidden bias in our case study are variables that are predictive of the health outcome that we have not considered in our analysis. The more liberal we are in including variables in our propensity score creation, the less chance that we have unaccounted hidden bias.

STEP 4: USE MATCHING TO CREATE TREATMENT/CONTROL GROUPS

Once the propensity scores have been created, they can be used to create treatment and control groups. In our case study we used "1 to 1 Greedy Matching". We matched each visit in our treatment group with a visit in our control group, based on the propensity score. With Greedy Matching, we first identify a 'caliper'. In the methodology presented in "Analysis of Observational Health Care Data Using SAS", a caliper is a defined width based on a proportion of the standard deviation of the logit of the propensity score (Faries, Leon, Haro, Obenchain, 2010). Each record from the treatment group is matched by propensity score to the nearest record in the control group that has not yet been matched. To be matched, records must have propensity scores that are within the caliper distance of each other.

Figure 2 below is a simplified example of how 1 to 1 matching works, with each treatment observation matched to a control observation based on the propensity score. In this scenario, we match each of our 5 treatment observations to a control observation. We have three left over control observations which are not used in the study.

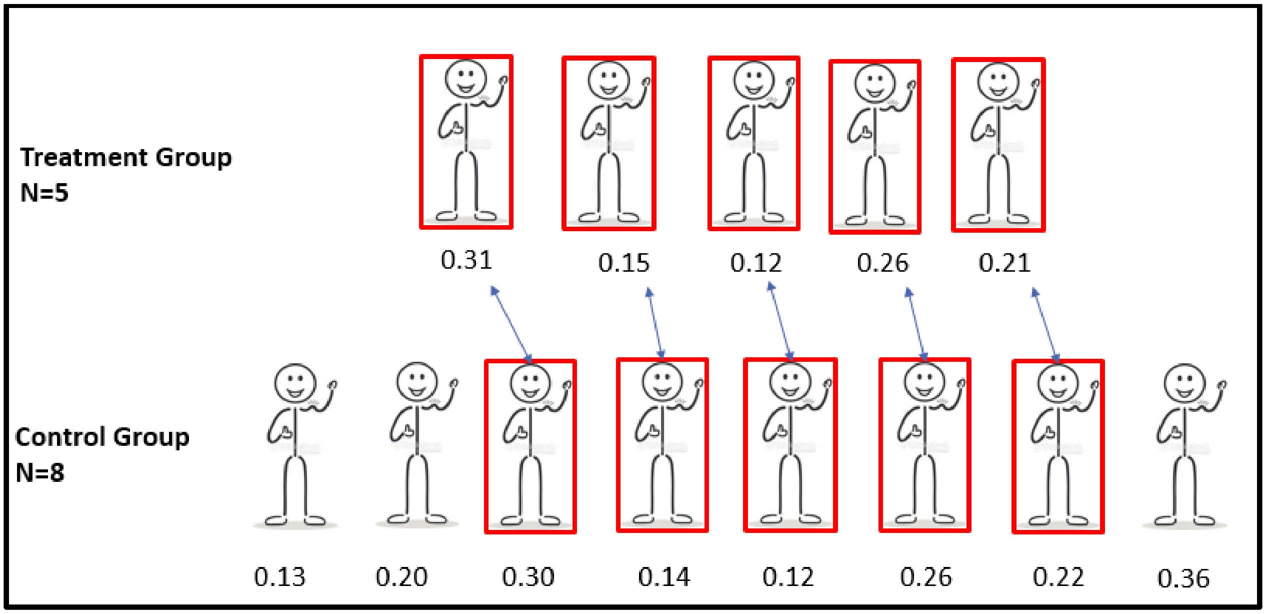
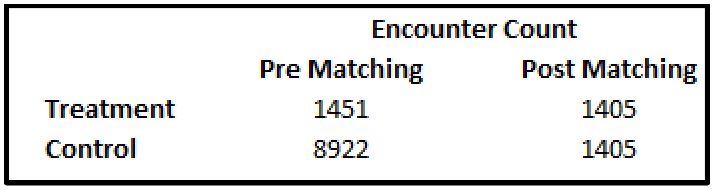


Figure 2. A simplified illustration of one to one matching based on propensity score

In our case study, we borrowed from code written by the Mayo Clinic that can be found in “Analysis of Observational Health Care Data Using SAS”. Specifically, we use the code from page 65 to the data step on page 67 that creates a dataset called “long”. This code creates our calipers and uses a macro called ‘%gmatch’ to perform one to one greedy matching. We recommend typing the code and running it step by step to obtain a better grasp of how it works. On page 66 the %gmatch macro names it’s parameters. If you have utilized the code thus far as directed in this paper, you will only need to update the following two parameters:

- data = ‘your dataset name’
- group = ‘your treatment variable name’

After we ran the code, we had a dataset named ‘long’, in which each record is represented by an encounter. It only contains encounters that are matched between the treatment and control groups. Encounters from the treatment or control groups that did not get matched are thrown out and not used. Table 5 shows that with one to one matching, we now have equal number of encounters in our treatment and control groups.



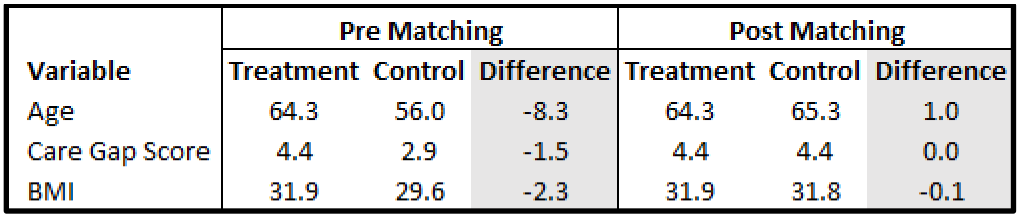
	Encounter Count	
	Pre Matching	Post Matching
Treatment	1451	1405
Control	8922	1405

Table 5. After propensity score matching, we have an equal number of encounters in our treatment and controls groups.

In the next step, we’ll see how successful we were in balancing our confounding variables through the propensity score matching we just performed.

STEP 5: TEST TREATMENT/CONTROL GROUP BALANCE

Once we created our treatment and control groups, we needed to test for balance among all the confounding variables that we identified. Table 6 shows that after one to one propensity score matching, the variables age, care gap score, and BMI are much more in balance between the treatment and control groups.



Variable	Pre Matching			Post Matching		
	Treatment	Control	Difference	Treatment	Control	Difference
Age	64.3	56.0	-8.3	64.3	65.3	1.0
Care Gap Score	4.4	2.9	-1.5	4.4	4.4	0.0
BMI	31.9	29.6	-2.3	31.9	31.8	-0.1

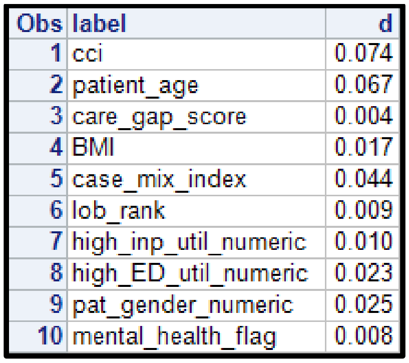
Table 6. Mean age, care gap score, and BMI before and after matching our treatment and control groups

We can utilize the code on page 68 to the middle of page 70 in "Analysis of Observation Health Care Data Using SAS®" to obtain a standardized difference between two means and determine if our confounding variables in our treatment and control groups are in balance. This code includes two macros: 1) 'cont' to test for the difference in means for continuous variables, and 2) 'binary', to test for the difference in means for binary variables. You will need to modify the code at the bottom of page 69, which calls the macros. Replace the parameters that are sent to the macro with your variable names. Here is the code we use to call the 'cont' and 'binary' macros, testing for the standardized difference in means between the control and treatment groups' patient age (continuous) and patient gender (binary) variables:

```
%cont(var=patient_age,label="patient_age");

%binary(var=pat_gender_numeric,label="pat_gender_numeric");
```

Table 7 gives the results of the standardized difference tests.



Obs	label	d
1	cci	0.074
2	patient_age	0.067
3	care_gap_score	0.004
4	BMI	0.017
5	case_mix_index	0.044
6	lob_rank	0.009
7	high_inp_util_numeric	0.010
8	high_ED_util_numeric	0.023
9	pat_gender_numeric	0.025
10	mental_health_flag	0.008

Table 7. The results of the standardized difference tests

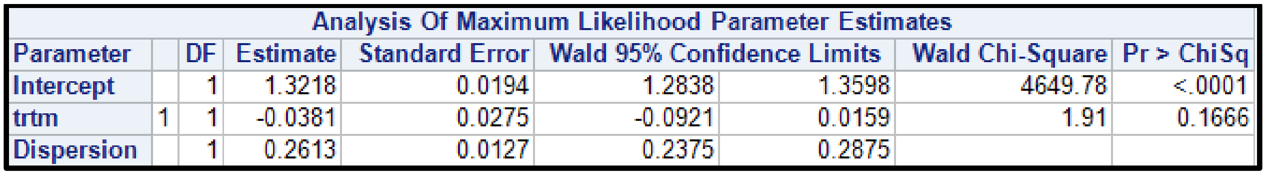
All observed confounding variables have means with relatively small standardized differences between the treatment and control groups. You will need to determine the level of standardized difference acceptable in your study. If you are left a with a strong confounder that has a high degree of non-balance after matching, you will need to research and perform a different matching methodology.

STEP 6: USE TREATMENT TO PREDICT THE OUTCOME

Now that you have your treatment and control groups and have checked your confounding variables for balance, you are ready to create a model. In creating a model with your treatment and control groups, you want to use a model that fits the distribution of your outcome variable. Because our health outcome had a negative binomial distribution, we created a negative binomial model with treatment (yes or no) as our predictor and our outcome as our dependent variable. Remember that the code you use to create your model should fit the data you're working with, so you will most likely have a different model and use different SAS code here:

```
proc genmod data = long;
  class trtm;
  model outcome = trtm/ type3 dist=negbin;
run;
```

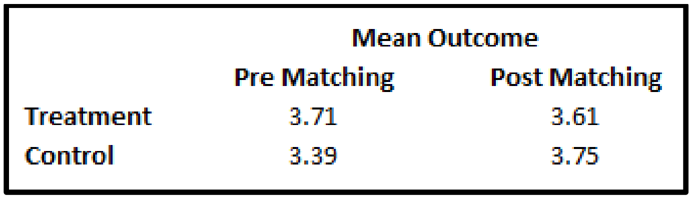
Once we balanced our confounding variables based on our propensity scores, a patient receiving treatment was no longer a significant predictor of our health outcome. Table 8 gives the results of our negative binomial model.



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.3218	0.0194	1.2838	1.3598	4649.78	<.0001
trtm	1	1	-0.0381	0.0275	-0.0921	0.0159	1.91	0.1666
Dispersion		1	0.2613	0.0127	0.2375	0.2875		

Table 8. The results of our model, with treatment as the predictor

Our p-value was 0.1666. Through propensity score matching we have shown that the treatment was not a significant predictor of our outcome value. Remember that before matching our treatment group had a significantly higher outcome value than our control group, indicating an adverse impact of treatment. Table 9 shows that after balancing for observed confounders, the outcome values for our treatment and control groups are more in line.



	Mean Outcome	
	Pre Matching	Post Matching
Treatment	3.71	3.61
Control	3.39	3.75

Table 9. Our Mean Outcome for our Treatment and Control Groups before and after matching

STEP 7: DOCUMENT AND VALIDATE

Steps 7-9 are perhaps the most important steps for answering analytical questions with propensity scores. It is likely that you will run several iterations of your code. For this reason it is highly recommended that anytime you save or share the results of your latest run, you archive your code that produced those results. Having a clear methodology to link your results to the code that produced them will help you understand why the results changed through different iterations. Any changes made to your initial model should be documented. If you share one set of results with your customer, and later follow up with a different set of results, you want to be able to explain what changed in the model and the reason the change was made. This level of clarity will not only improve your final product, but it will give your work transparency and help improve the credibility of your analysis with your customer.

In this step you should also validate your work. You may find mistakes in the source data, errors you made when setting up your data, or adjustments that need to be made to your code that require you to go back to an earlier step in the process.

STEP 8: STATISTICAL APPROACH PEER REVIEW

As mentioned in the abstract, this paper is for both those attendees who have never used propensity scores and those who have a basic understanding of propensity scores but are unsure how to begin using them in SAS. One of the most efficient ways to learn a new skill in SAS is to jump into the subject matter at hand. To reach a higher level of confidence in your work, we highly recommend that you have your work peer reviewed, regardless of whether you are a seasoned statistician or someone with little statistical background. There are many opportunities to make mistakes that can impact your conclusions. Through peer review you can catch these mistakes before sharing your findings. Having your work peer reviewed by a colleague at your place of employment, a professor at a local university, or any other person with experience using propensity scores will result in a better analysis and provide a great learning opportunity for you.

In the Kaiser Northwest region, we have a “Center for Health Research”, which has many seasoned statisticians with experience using propensity scores. We reviewed our analysis with them and were provided valuable feedback that helped improve our analysis. This gave us more confidence in our results.

STEP 9: CUSTOMER REVIEW

In Step 1 we identified the importance of conducting thorough interviews with customers and subject matter experts with intimate contextual knowledge of the question, as it reduces the number of times your analysis will need to be reworked. This analytical work is best performed when regularly consulting with your customers, and should undergo a final review. Before you present your results to your customers, you should make sure they agree with the business logic incorporated into the analysis. For example, your customer may have insight into whether:

- outliers should be included
- important predictors have been left out
- pertinent observations are included and the correct filters are placed on the data
- changes occurred in operational processes during the time the observational data was captured that impact how the data is reflected in your database

Once this review is complete, it's possible you will need to return to an earlier step to rework your model with the new information you have obtained.

STEP 10: PRESENT RESULTS

Upon completion of the analysis, prepare to communicate your work in a way that best meets your customer's needs and expectations. Is the customer expecting a high-level presentation or a paper with statistical details included? The better you understand your customer's statistical background and knowledge, the more impactful your analysis will be. Providing too much statistical detail to a customer with no statistical background can deflect from the primary conclusions of the analysis, while only serving to confuse. Glossing over key details with customers with more statistical knowledge can leave them questioning the analysis.

Also, it's important to understand what your customer will do with the results. Your customer may only be interested in the answer to the analytical question. Or they may want to perform additional analysis with the data you provide. A power point presentation may be sufficient for a customer only interested in high level answers. Customers interested in performing additional analysis would also require data in a format which allows them to utilize the data, such as an excel spreadsheet.

Finally, whatever format you decide on for the communication of your analysis, take the time to ensure it is presented in a clear and professional manner. It is likely you've spent many hours on your analysis, and the communication of your work should reflect that. A technically sound analysis that is presented poorly may lead your customers to question its credibility. For more formal presentations, a final peer review in which you receive feedback is always helpful and can help boost your confidence that your product is ready for delivery to your customer.

CONCLUSION

In our case study, we learned that patients receiving treatment had a statistically significant higher mean outcome value than those patients not receiving treatment. If we ended our analysis at that point and provided the results to our customers, we would have misled them and likely caused them to make decisions based on false assumptions. Fortunately, we understood that our study was not a randomized controlled trial. The patients in our treatment and control groups were not selected randomly. Because of this we had to deal with selection bias and confounding variables. Propensity score one to one matching allowed us to balance confounding variables between our treatment and control groups. We felt comfortable that we were comparing two groups with similar characteristics, in which the primary difference between the two groups was that one group received the treatment and the other group did not. Once our groups were balanced, we learned that there was not a significant difference in our outcome. Several layers of review and a presentation of the results with the customer needs in mind resulted in a satisfied customer that could make decisions based on a sound analysis.

REFERENCES

Faries, D., Leon, A., Haro, J., Obenchain, R. 2010. *Analysis of Observational Health Care Data Using SAS®*. Cary, NC: SAS Institute Inc.

Westreich, D., Cole, S., Funk, M., Brookhart, M., Sturmer, T. 2011. “The role of the *c*-statistic in variable selection for propensity score models.” *PubMed Cental,* 20(3): 317–320.

RECOMMENDED READING

- *Analysis of Observational Health Care Data Using SAS®*

CONTACT INFORMATION

Thomas Gant
Data & Information Management Enhancement (DIME)
Kaiser Permanente
503-813-4906
thomas.e.gant@kp.org

Keith Crowland
Data & Information Management Enhancement (DIME)
Kaiser Permanente
503-813-4089
keith.d.crowland@kp.org