PharmaSUG 2016 – Paper 36

What's the Case? Applying Different Methods of Conducting Retrospective Case/Control Experiments in Pharmacy Analytics

Aran Canes, Cigna, Bloomfield, CT

ABSTRACT

Retrospective Case/Control matching is an increasingly popular approach to evaluating the effect of a given treatment. There is some theoretical literature comparing different methods of case/control matching, but there is a lack of empirical work in which each of these methods is employed. In this paper, I use SAS® to conduct a retrospective case/control experiment on the efficacy and safety of two distinct drug options using Propensity Score Matching, Mahalanobis Metric Matching with Propensity Score Caliper and Coarsened Exact Matching. I then compare outcomes from all three methods to try to develop a sense of the advantages and disadvantages of each. In this example, despite a considerable lack of overlap in the output datasets, all three methods lead to similar results: evidence of increased effectiveness for one of the drugs with little to no difference in safety. More generally, when a researcher needs to choose one method over another, I conclude that the choice should be guided by their understanding of how close the pairwise distance needs to be between cases and controls and the degree to which it is appropriate to measure the treatment on only a part of the pre-matched case population.

Key words: Retrospective Analyses, Propensity Score Matching, Mahalanobis Metric Matching, Coarsened Exact Matching

INTRODUCTION

With the growth in modern computational power, the conduct of retrospective studies by creating similar case and control populations across known covariates is becoming more frequent. These observational studies are particularly pertinent to the pharmaceutical industry as cost and feasibility issues often prevent the use of a double-blind clinical trial. Insurance companies and Medicaid/Medicare agencies may, for example, want to test the efficacy of two distinct drugs with the same indication but cannot randomly assign customers into cases and controls. Instead, these entities can try to simulate the rigor of a clinical trial by looking retrospectively at the claims data of customers who used these medications, creating matched case/control groups and then evaluating outcomes on these matched populations. Given this increasing use, there has been a corresponding growth in the choice of methods available to researchers to create matched case and control populations. Prominent among them are Propensity Score Matching, Mahalanobis Metric Matching with Propensity Score Caliper and Coarsened Exact Matching. There have been some theoretical comparisons made of these different methods[1] but very little published material comparing them in practice. This paper reviews the similarities and differences between these three methods in a real world setting, with guidance given as to the advantages and disadvantages of each, particularly in regard to pharmacy analytics. All analyses were performed using SAS® Enterprise Guide 6.1. I will first briefly review the history and theory of all case/control studies in order to provide an appropriate context. Then I will review the methods and present the background of a particular example. Next, I will look at the differences in the matched case/control populations and outcomes for each approach. Finally, I will make some practical suggestions for the pharmaceutical researcher on when they might want to prefer one method over another.

HISTORY AND THEORY OF CASE/CONTROL STUDIES

Literature discussing matching methods appears all the way back to the 1940s but the theory was not given a rigorous theoretical basis until the early 1970s. In 1973, Donald Rubin described the primary purpose of all case/control analyses as discovering the Average Treatment Effect (ATE). This is defined as the average difference in outcomes associated with undergoing a particular treatment. Since each individual either does or does not receive the treatment we cannot directly observe the ATE for that individual. Instead, we can only compare individuals who received the treatment with those who did not. Creating a population who did not receive the treatment that is similar to the population who did is then a central goal of all retrospective case/control studies. Mathematically, the treatment may be defined for individual Y as Y(1)-Y(0) where 1 stands for receiving the treatment. If we have a vector of p covariates denoted by X for each individual then the treatment effect can be redefined as E(Y(1)|X)- E(Y(0)|X). Here the expectation given the covariates signifies that the treatment and control populations are balanced on all variables which could have an effect on the outcome of interest. From the difference in outcomes between the matched treated and untreated populations one can then calculate the Average Treatment Effect. The major premise of all case/control matching is the ignorability assumption, which specifies that the assignment to case or control is independent of all outcomes given the covariates. If a variable is left out which was significant in determining whether an individual was case or control, and that variable has an effect on the outcome(s), the researcher is introducing

bias into the study since the effect of the treatment cannot now be separated from the effect of the left out covariate. Rubin has further formalized case/control matching by classifying them as one particular variation of the analysis of treatment effect. The difference then between retrospective case/control matching and randomized experiments is simply in the assignment mechanism. In randomized trials participants are sorted into case and controls randomly while in the retrospective study certain variables were used to sort participants into case and controls. The central task of observational studies is then to recover the mechanism, i.e. variables, which were used to sort out the participants.[2]

THREE METHODS FOR CONDUCTING OBSERVATIONAL STUDIES

The most widely used method for conducting retrospective case/control studies is propensity score matching (PSM). Developed by Rubin and Rosenbaum in 1983[3], PSM calculates the propensity of each individual to be in either the case or control population based on the observed covariates. This propensity is given by performing a logistic regression where the dependent variable is a case/control indicator and the predictor variables are the covariates influencing which category the individual fell into. In this stage of PSM, the outcome data is withheld to avoid researcher bias. Instead, one conducts the analysis strictly to retrospectively determine the assignment mechanism which resulted in the treated/untreated populations for which there is retrospective data. Once each individual has a calculated propensity to be in either the case or control groups, either the predicted probability or the actual logit, a matching algorithm is employed which matches individuals with similar propensities. Outcomes are then evaluated on the matched population. Mahalanobis matching (MNM) with Propensity Score Calipers refines PSM. At the 2006 PharmaSUG conference, Feng, Jun and Xu[4] presented this method of first computing the traditional propensity score and then using a distance measure to match individuals both by their propensity score and the distance the individuals' covariates are from one another. More simply, the macro computes the Mahalanobis distance, a generalization of the Euclidean distance which takes into account the covariance between the different variables, between all case and controls within a certain threshold of the propensity score. The Propensity Score caliper is used to make certain that the distance is sufficiently small to be considered a good match and to reduce the total number of pairs the Mahalanobis distance must be calculated between. The purpose of this method is to reduce the absolute distance between each case/control pair compared to traditional PSM. Coarsened Exact Matching (CEM) provides a more radical approach to reducing the distance between case/control pairs.[5] CEM allows the researcher to determine the appropriate granularity, or coarseness, for a match across all covariates. For example, if we have a continuous income variable the researcher can create low, middle and upper income levels and then match exactly on the coarsened variables. The rationale is that a researcher sufficiently familiar with the covariates should know how close a match is necessary on each covariate to sufficiently control for these confounders while still allowing enough slack to produce sufficient matches to evaluate the treatment effect.

CASE STUDY-EFFICACY AND SAFETY OF DRUG A VERSUS DRUG B

Cigna, one of the five largest health care insurance companies, often has to make decisions based on the effects of different drugs on consumers. While one can use clinical trial findings on these drugs to guide these decisions, insurance companies are increasingly relying upon their own patient outcomes stored in their data warehouses. In this instance, given that Drug A has been available for over two years, Cigna would like to leverage our own internal data to see whether we are observing the increased efficacy and safety for Drug A versus the more traditional Drug B which was seen in the clinical trials. However, the characteristics of the population who uses Drug A may be very different from the Drug B population. Because of this, I used Propensity Score Matching, Mahalanobis Metric Matching with Propensity Score Caliper and Coarsened Exact Matching to create populations of similar customers on which to evaluate the efficacy and safety of Drug A.

RESULTS

The first stage in either PSM, MNM or CEM is to identify the covariates which function as the assignment mechanism in determining whether a customer was case (Drug A user) or control (Drug B user). Based on our prior knowledge of factors determining when a doctor is likely to prescribe one drug or the other we identified five covariates: Age, Region, Gender, Presence of Cardio-Ablation and Retrospective Risk Number. Retrospective risk is a score based on prior year diagnoses which measures the overall health of the patient. Statistically significant differences for these confounders can be assessed by performing chi-square tests on categorical variables and t-tests on continuous variables as in the following examples:

```
proc freq data=prematchedpopulation;
    table drugtype*agecat /chisq nopercent nocol;
    title 'UNMATCHED DATA';
run;
```

```
proc ttest data=prematchedpopulation;
          class drugtype;
          var retrsp_risk_num;
          title 'UNMATCHED DATA';
run;
```

Table 1 below has the pre-matched distribution of these variables for both Drug A and Drug B. The Drug B population comes from different parts of the country, has a higher probability of being female, has fewer patients with a cardio-ablation and is considerably sicker.

Confounding Variables		Drug A	Drug B	P-Value
Number of Customers		194	518	
Age	< 17	0%	0%	0.6791
	18 to 24	0%	0%	
	25 to 34	1%	1%	
	35 to 44	3%	4%	
	45 to 64	58%	53%	
	65 or Older	38%	41%	
Region	Midwest	10%	20%	<0.0001
	Northeast	13%	6%	
	Other	24%	30%	
	South	44%	31%	
	West	9%	13%	
Gender	Females	24%	32%	0.0303
	Males	76%	68%	
Cardio-Ablation Procedure		10%	3%	0.0007
Mean Retrospective Risk Score		4.90	6.69	0.0002

Table 1 Pre-Matched Population Characteristics

The next stage in all of these methods is to create a matched population of case and controls. In PSM, one first uses a logistic regression to model the propensity to be in either the case or control population as in the following example:

```
proc logistic data=prematchedpopulation;
       class agecat gender region ablation;
       model drugtype (event='1') = agecat region gender retrsp_risk_num ablation;
       score data=prematchedpopulation out=prematchedpopulation2;
run;
```

One then employs an algorithm for matching case and control customers based on this predicted propensity. In this case I used a greedy matching algorithm[6] from a previous SAS® Users Group Conference. One can find the code in the source provided in the references.

Table 2 below shows the post-matching population characteristics for Propensity Score Matching.

Confounding Variables		Drug A	Drug B	P-Value
Number of Customers		189	189	
Age	< 17	0%	0%	0.2765
	18 to 24	0%	0%	
	25 to 34	1%	0%	

Confounding Variables		Drug A	Drug B	P-Value
	35 to 44	3%	2%	
	45 to 64	58%	53%	
	65 or Older	38%	45%	
Region	Midwest	11%	10%	0.8806
	Northeast	13%	11%	
	Other	25%	29%	
	South	43%	40%	
	West	9%	11%	
Gender	Females	24%	20%	0.3223
	Males	76%	80%	
Cardio-Ablation Procedure		8%	7%	0.6945
Mean Retrospective Risk Score		4.98	4.46	0.2014

Table 2 Post-Matching Differences PSM

PSM was able to 189 out of the 194 customers in the Drug A population to a Drug B member. All of the covariates now have statistically insignificant differences between the Drug B and Drug A population. However, unlike exact matching, some differences remain. In MNM, one first uses the propensity score to create a caliper, then matches on the closest Mahalanobis distance. To perform this, I used a macro presented at a previous SAS® Pharmacy Users Group Conference[4]. The paper cited in the references contains the code for this macro. Table 3 below shows the post-matched covariate statistics for MNM.

Confounding Variables		Drug A	Drug B	P-Value
Number of Customers		190	190	
Age	< 17	0%	0%	0.8540
	18 to 24	0%	0%	
	25 to 34	1%	2%	
	35 to 44	3%	4%	
	45 to 64	58%	54%	
	65 or Older	38%	41%	
Region	Midwest	10%	11%	0.6251
	Northeast	12%	11%	
	Other	25%	31%	
	South	44%	41%	
	West	9%	6%	
Gender	Females	24%	23%	0.7163
	Males	76%	77%	
Cardio-Ablation Procedure		8%	8%	1
Mean Retrospective Risk Score		4.96	4.74	0.5840

Table 3 Post-Matching Differences MNM

Like PSM, MNM has created a matched case and control population with statistically insignificant differences across the covariates. 190 out of the 194 Drug A customers were matched, an improvement of one over PSM. The last approach to consider is Coarsened Exact Matching. In CEM the variables are first coarsened to the amount of

difference allowed by the researcher and then an exact match is performed on the coarsened variables. For the Drug A/Drug B comparison I left the categorical variables as they were and employed Sturges Rule (Number of Bins=$\log_2$(Number of Observations)+1) to create bins for the one continuous variable (Retrospective Risk Number). A macro which performs Coarsened Exact Matching may be found in the link cited in the references.[7] Table 4 below shows the post-match statistics on the covariates.

Confounding Variables		Drug A	Drug B	P-Value
Number of Customers		167	167	
Age	< 17	0%	0%	1
	18 to 24	0%	0%	
	25 to 34	1%	1%	
	35 to 44	2%	2%	
	45 to 64	59%	59%	
	65 or Older	38%	38%	
Region	Midwest	10%	10%	1
	Northeast	10%	10%	
	Other	26%	26%	
	South	47%	47%	
	West	8%	8%	
Gender	Females	24%	24%	1
	Males	76%	76%	
Cardio-Ablation Procedure		3%	3%	1
Mean Retrospective Risk Score		4.57	4.66	0.8343

Table 4 Post-Matching Differences CEM

Because this is an exact match the categorical variables are now identical. However, there has been a cost. Only 86% of the Drug A customers have been matched to Drug B customers. Prior to looking at the different outcomes of these three methods I calculated the mean Mahalanobis distance for each methodology. Table 5 below shows this distance:

Method	N	Mean Mahalanobis Distance
Propensity Score Matching	189	3.26
Mahalanobis Matching with Propensity Score Caliper	190	2.62
Coarsened Exact Matching	167	0.30

Table 5 Mean Mahalanobis Distance

As one can see, there is a dramatic difference in the pairwise Mahalanobis distance between CEM and the other two methods. An analysis of variance confirmed that this result was highly significant (<0.0001). By balancing the populations on all variables simultaneously, PSM and MNM lead to a significantly greater difference across all covariates for each case and control. However, the dramatically reduced pairwise distance between each case and control provided by CEM has come with the cost of not matching more than 20 cases. Not only does this lead to less statistical power, we are now evaluating the average treatment effect on a different Drug A population than in the other two instances. Interestingly, there was only about 50% overlap between the Drug B populations between each method. Table 6 below shows the overlap between each of the control populations:

Method	N	Overlap PSM	Overlap MNM	Overlap CEM

Propensity Score Matching	189	N/A	91(48%)	86(45%)
Mahalanobis Matching with Propensity Score Caliper	190	91(48%)	N/A	80(42%)
Coarsened Exact Matching	167	86(51%)	80(48%)	N/A

Table 6 Overlap

Because of the differences in the control population between these methods it would not be surprising if there are differences in the clinical outcomes: the number of customers who had a condition related negative event and the number of customers who experienced side effects. Table 7 below shows the different outcomes associated with each of the three methods.

Drug	Matched Population	Number of Members with Side Effects	Percentage of Members with Side Effects	P-Value (Side Effects)	Number of Members with Adverse Outcome	Percentage of Members with Adverse Outcome	P-Value (Adverse Outcome)
Propensity Score Matching							
Drug A	189	19	10.05%	0.5945	4	2.11%	0.0707
Drug B	189	16	8.47%		12	6.35%	
Mahalanobis Metric Matching							
Drug A	190	19	10.00%	0.1896	4	2.11%	0.1714
Drug B	190	12	6.32%		10	5.25%	
Coarsened Exact Matching							
Drug A	167	18	10.78%	0.2509	2	1.20%	0.1041
Drug B	167	12	7.19%		8	4.79%	

Table 7 Outcomes

All three methods had similar side effect results: a slight, statistically insignificant increase in the number of customers experiencing side effects for Drug A customers. Similarly, each of the methods provides a dramatic difference in the efficacy of Drug A over Drug B. While the rarity of adverse results in the overall population prevents this outcome from being statistically significant at 0.05 level, the clinical trial literature appears to be confirmed: Drug A significantly reduces the probability of the condition specific negative event. At this point the researcher must look at the reasons CEM matched only 167 out of the 194 case customers (Drug A) and determine whether this smaller population is a more accurate measure of the Average Treatment Effect than PSM or MNM. Analysis reveals that exact matching on region and whether an ablation procedure was performed are the primary reasons these 27 customers were not matched. The researcher at this point has to either accept this reduced population or rerun CEM with further coarsening of region. Since the larger population produced by PSM and MNM reported similar results, I did not see the need to perform CEM again with coarser variables.

PRACTICAL ADVICE FOR THE PHARMACEUTICAL RESEARCHER ON CHOICE OF METHOD

In this instance all three methods reported similar outcomes. However, even given these similar results we can still identify of some advantages and disadvantage to each method. Propensity Score Matching allows one to easily control for a large number of covariates and leads to matched case/control pairs with insignificant population wide differences. It also tends to retain a large percentage of the pre-matched case population in the post-matched population. However, this is achieved at the cost of large pairwise differences between each case and control. Since all that is matched is the propensity score given all predictor variables combined, each case/control pair can have large differences in each of the individual covariates. If a researcher's understanding of the retrospective data is that minimizing pairwise distance, as opposed to population wide distance, is important this is not the preferred alternative. Mahalanobis metric matching with propensity score caliper improves on the pairwise match between each case and control. In this instance, the choice of the propensity score caliper actually led to more overall matches than

PSM. However, MNM is computationally intensive as the Mahalanobis distance has to be calculated between many case and control combinations. In addition, despite minimizing the pairwise distance between each case and control within a certain caliper, the average pairwise distance is an order of magnitude larger than CEM. Coarsened Exact Matching allows the researcher to determine in advance how coarsened the match on each variable should be. This allows for more control over the post-matched population. Because CEM is an exact matching algorithm, all statistics (not only the mean) in the post-matched populations are balanced. There is also no need to check post-matching for a random distribution across the covariates. However, if one does not know the appropriate coarsening for each covariate in advance CEM there is no distinct advantage to CEM over exact matching. CEM can also lead to post-match sample sizes that are smaller than other approaches, thus decreasing the statistical power of a test. More importantly, since the average treatment effect is a metric calculated on a given population, CEM may lead to different outcomes than methods which lead to more matches since CEM measures the ATE on a different population.

CONCLUSION

The contemporary researcher has a variety of methodological options available to conduct retrospective case/control studies. This paper has looked at three of the most frequently chosen: Propensity Score Matching, Mahalanobis Metric Matching and Coarsened Exact Matching. In this instance, the methods returned the same primary outcomes even though different matches were made in the control population. Although there is no one rule to guide the choice of method, calculating the Mahalanobis distance suggests that CEM produces closer pairwise matches than PSM or MNM. The cost of these differences is a significantly smaller post-matched population than in the other instances. A researcher should choose a method based on their understanding of the underlying data and the desire to balance the need of small differences between each case/control pair with the need to fully represent the case population in the post-matched data and resulting outcomes.

REFERENCES

1. Stuart, Elizabeth. February, 2010. “Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Sciences.” *Statistical Science*.1-21. Bethesda, MD: Institute of Mathematical Statistics.

   King, Gary, Nielsen Richard, Coberley, Carter and Pope, James. “Comparative Effectiveness of Matching Methods for Causal Inference.” December, 2011. Available at http://gking.harvard.edu/publications/comparative-effectiveness-matching-methods-causal-inference

   Ho, Daniel, Imai, Kosuke, King Gary and Stuart, Elizabeth. January, 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.” *Political Analysis*. 199-236. New York, NY. Oxford University Press.
2. Rubin, Donald B. 2008. “For Objective Causal Inference Design Trumps Analysis.” *The Annals of Applied Statistics*. 808-840. Bethesda, MD: Institute of Mathematical Statistics.
3. Rosenbaum, Paul R. and Rubin, Donald B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika*. 41-55. New York, NY. Oxford University Press.
4. Feng, Wu Wei, Jun, Yu and Xu, Rong. 2006. “A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study.” *Proceedings of the Pharmacy SAS Users Group.* Available at http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf
5. Iacus, Stefano M, King, Gary and Porro, Giuseppe. 2011. “Causal Inference Without Balance Checking: Coarsened Exact Matching,” Political Analysis. *Political Analysis*. 1-24. New York, NY: Oxford University Press.
6. Parsons, Lori S. 2004. “Performing a 1:N Case Control Match on Propensity Score.” *Proceedings of the SAS Users Group International Conference*. Available at: http://www2.sas.com/proceedings/sugi29/165-29.pdf
7. King, Gary. “CEM: Coarsened Exact Matching Software. 2014.” Available at: http://gking.harvard.edu/cem

ACKNOWLEDGMENTS

The author would like to thank Dr. Michael Manocchia, Dr. Saad Aslam and Mr. Jigar Shah for providing me the opportunity to conduct this analysis and for helpful suggestions throughout.

CONTACT INFORMATION

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

Name: Aran Canes

Enterprise: Cigna
Address: 900 Cottage Grove Rd.
City, State ZIP: Bloomfield, CT 06002
Work Phone: 860-226-4443
E-mail:aran.canes@cigna.com