

Bias Reduction in Observational Cohort Studies - obtaining matched cohorts using propensity scores

Shelley Fordred, Savvy Stats Ltd, Heston, UK

Victor Kiri, FV and JK Consulting, Surrey, UK

Knut Mueller, UCB Biosciences, Monheim

ABSTRACT

Matched cohort designs where different treatments are compared offer a lot of challenges with regard to the selection of variables for matching. One popular method of getting around the issues associated with matching on several individual variables, especially the associated problem of insufficient study sample size is to use estimated propensity scores to obtain matched cohorts that are comparable. By way of illustration of the approach from the programmer's perspective, this paper describes the steps which our recent experience on real-life observational data suggests as vital in the estimation of propensity scores for such purpose.

Keywords: Propensity Scores, Confounding, PROC LOGISTIC, Matching

INTRODUCTION

Cohort studies involve following subjects who have been exposed to a certain treatment such as a drug. Matching is where the exposed subjects are matched to those subjects not exposed to the treatment on certain covariates which are considered necessary to facilitate comparability between the two groups⁴. Matching is done to reduce confounding in cohort designs, for example the exposed subjects may be older than the unexposed and so we match subjects with comparable age in both cohorts. The aim is to compare subjects with similar baseline characteristics which the randomization routine facilitates in randomised controlled studies. The variables used for matching are normally those which do not vary over time (i.e. constant) - largely demographic variables such as height, weight, disease history, etc.

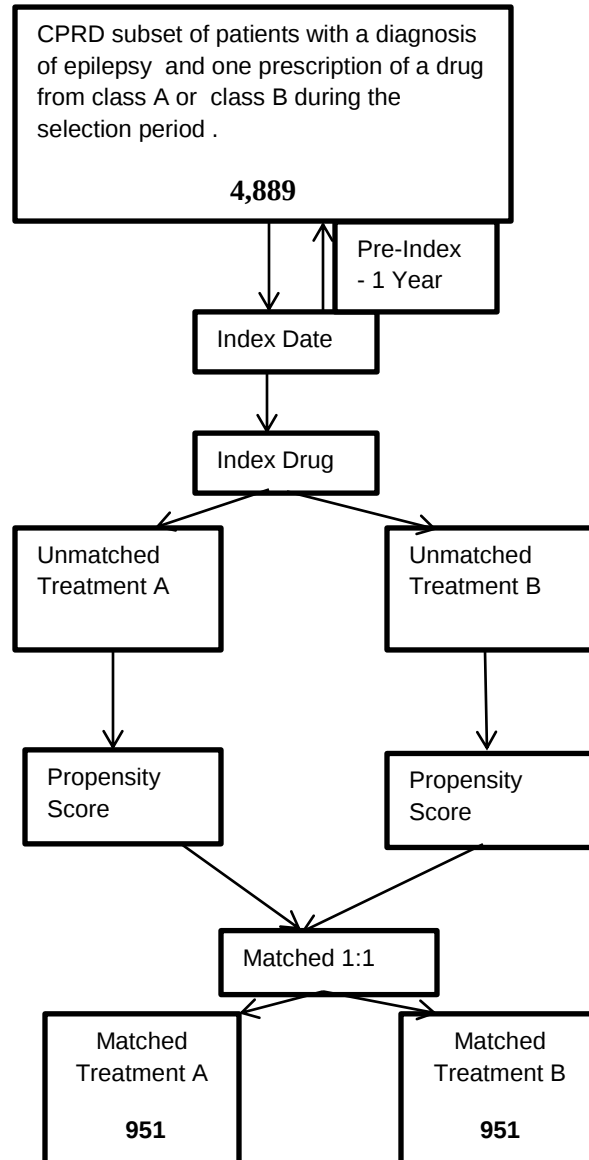
The limitations of matching includes the need for a large sample size due to the inevitable discarding of unmatched patients and identification of suitable/appropriate variables for matching from a list of potential confounders that can potentially run into thousands. However, when matching is carried out correctly, confounding can be significantly reduced when compared with the estimates obtained from the use of traditional regression methods^{5,1,4}.

As an alternative to matching on individual variables, the two treatment groups can be matched on composite scores such as the propensity scores (PS)². A propensity score is the likelihood of a patient being exposed to a particular treatment or intervention given a set of explanatory variables.. The scores are usually estimated from a suitable statistical model that involves individual patient-level variables as covariates. The subjects are subsequently matched on their propensity score. Subjects from both treatment groups are matched according to their probability of getting , for example treatment B, based on the covariates that went into the propensity model.

One of the main challenges of the propensity scores methodology is on the choice of a suitable matching routine for the particular study³. The list of routines includes (a) caliper, (b) closest available neighbour, (c) nearest available Mahalanobis metric and (d) radius matching^{6,1}. The other area of choice is whether the unexposed patients can be selected more than once for matching (i.e. with or without replacement).

This paper will describe the programmers experience in a recent study- the selected method of matching, detail description of its implementation, the challenges we encountered and how these were resolved.

STUDY DESIGN



The aim of this study was to compare the direct healthcare cost of treating epileptic patients on a particular class of drugs from treatment class A against another class of drugs from treatment B using the propensity scores-matched cohort methodology on a 1:1 basis.

The study patients must have been diagnosed with epilepsy and must also have had at least one prescription of a drug from either treatment class A or B during the selection period 01/01/2001 to 31/12/2010. 4,889 subjects were eligible for this study. The index date was defined as the start date of the first use of a drug from either treatment A or B during the selection period, here referred to as the index drug and thus marked the date of cohort entry into the respective treatment group. The pre-index (baseline) period started 1 year prior to the index date.

The propensity score was defined as the probability of the patient being prescribed treatment B and the scores estimated in SAS® using a multivariate logistic regression model with PROC LOGISTIC.

The covariates for the propensity score model were all based on the results of an exploratory analysis on the characteristics of the patients at their index date as well as guidance from the literature about the treatments for the condition. The initial list included those about general health and cost related characteristics such as age, gender, smoking status and number of general practice (GP) visits. Others included those specifically related to epilepsy such as comorbidities, medications as well as non-epileptic medications. As part of the exploratory process, only variables on which the two treatment groups were significantly different or those with an incidence of 1% or more in either treatment group were included in the propensity scores estimation model. The adequacy of the estimated propensity scores for the purpose of achieving comparable matched cohorts was also assessed by testing for differences between the two groups on the individual covariates.

Matching of the cohorts was performed using a validated macro which applied the caliper matching method without replacement. The caliper method defines a range of values in which the propensity score values of both cohorts have to fall and is usually defined in units of the standard deviation of the propensity score. The patient from treatment B is then matched to a patient from treatment A whose propensity score value is the closest to their propensity score.

PROPENSITY SCORE CREATION

The propensity scores were derived using a multivariate logistic regression model using PROC LOGISTIC. Before the PROC LOGISTIC code could be developed, the covariates for the propensity score model were selected from the analysis of the unmatched characteristics at index date and the pre-index period (baseline):

SELECTION OF THE BASELINE CHARACTERISTICS

The treatment differences for the unmatched baseline characteristics were produced using the t-test for the continuous variables and the chi-square test for the categorical variables (reference Table 1). All of the baseline characteristics i.e. age (subjects in treatment A were substantially older than in treatment B), age category, patient's gender (subjects in treatment A were more likely to be male), smoking status, body mass and the time period were statistically significant with a p-value of <0.05 and so were selected as covariates for the propensity score model. For the epilepsy related clinical variables time since first epilepsy diagnosis to the index date, time since most recent epilepsy diagnosis to the index date, first epilepsy diagnosis on index date, index drug as first ever drug for treating epilepsy, epilepsy type, history of status of epilepticus and Germaine-Smith epilepsy specific comorbidity index over the pre-index period were all statistically significant with a p-value of < 0.05 and so these variables were also included in the propensity score model.

Table 1 - Statistically significant baseline covariates before and after matching

Variable - Category	Statistic	TMT A (N=2752) Unmatched	TMT B (N=2137) Unmatched	P-value Unmatched	TMT A (N=951) Matched	TMT B (N=951) Matched	P-value Matched
Age at index date (Years)	N	2,752	2,137		951	951	
	Mean (Std)	54.2 (19.8)	45.9 (19.9)	<.0001	47.7 (19.9)	48.0 (20.0)	0.7969
Age Band				<.0001			1.0000
16-20	N (%)	116 (4.2%)	239 (11.2%)		75 (7.9%)	75 (7.9%)	
21-30	N (%)	261 (9.5%)	349 (16.3%)		144 (15.1%)	144 (15.1%)	
31-40	N (%)	404 (14.7%)	344 (16.1%)		167 (17.6%)	167 (17.6%)	
41-50	N (%)	434 (15.8%)	380 (17.8%)		160 (16.8%)	160 (16.8%)	
51-60	N (%)	406 (14.8%)	285 (13.3%)		133 (14.0%)	133 (14.0%)	
61-70	N (%)	425 (15.4%)	236 (11.0%)		119 (12.5%)	119 (12.5%)	
71-80	N (%)	434 (15.8%)	191 (8.9%)		94 (9.9%)	94 (9.9%)	
81-90	N (%)	237 (8.6%)	96 (4.5%)		47 (4.9%)	47 (4.9%)	
91+	N (%)	35 (1.3%)	17 (0.8%)		12 (1.3%)	12 (1.3%)	
Patient's Gender				<.0001			0.7130
Female	N (%)	1,243 (45.2%)	1,339 (62.7%)		506 (53.2%)	514 (54.0%)	
Male	N (%)	1,509 (54.8%)	798 (37.3%)		445 (46.8%)	437 (46.0%)	
Smoking status				<.0001			0.7035
Non-Smoker	N (%)	2,388 (86.8%)	1,749 (81.8%)		807 (84.9%)	801 (84.2%)	
Smoker	N (%)	364 (13.2%)	388 (18.2%)		144 (15.1%)	150 (15.8%)	
Index date calendar year				<.0001			0.0669
2001	N (%)	298 (10.8%)	70 (3.3%)		45 (4.7%)	55 (5.8%)	
2002	N (%)	326 (11.8%)	88 (4.1%)		53 (5.6%)	63 (6.6%)	
2003	N (%)	312 (11.3%)	121 (5.7%)		86 (9.0%)	72 (7.6%)	
2004	N (%)	402 (14.6%)	147 (6.9%)		109 (11.5%)	85 (8.9%)	
2005	N (%)	314 (11.4%)	190 (8.9%)		106 (11.1%)	96 (10.1%)	
2006	N (%)	275 (10.0%)	222 (10.4%)		121 (12.7%)	109 (11.5%)	
2007	N (%)	256 (9.3%)	255 (11.9%)		121 (12.7%)	119 (12.5%)	
2008	N (%)	225 (8.2%)	293 (13.7%)		106 (11.1%)	123 (12.9%)	
2009	N (%)	206 (7.5%)	371 (17.4%)		130 (13.7%)	119 (12.5%)	
2010	N (%)	138 (5.0%)	380 (17.8%)		74 (7.8%)	110 (11.6%)	

SELECTION OF THE CO-MEDICATION AND COMORBIDITIES

The treatment differences and incidences for the co-medication and comorbidities were determined during the pre-index period using the chi-square method. The variables were selected if either of the treatment differences were statistically different with a p-value of <0.05 or if the comorbidity or the drug had an incidence of >1% in either treatment. For example, chemotherapy was not selected as a variable for the propensity score model as the treatment differences were not statistically different and its incidence was less than 1%, but hormonal contraceptives were included as the treatment difference was statistically significant for this class of drug (Table 2). The co-medication and comorbid conditions were each coded as a dichotomous variable defining the presence or absence of a comorbidity or a co-medication.

Altogether 318 variables relating to the baseline characteristics, co-medication and comorbid conditions were selected for the model,

Table 2 - Statistically significant baseline co-medication before and after matching

Variable		TRT A (N=2752)	TRT B (N=2137)	P-value	TRT A (N=951)	TRT B (N=951)	P-value
- Category	Statistic	Unmatched	Unmatched	Unmatched	Matched	Matched	Matched
Hormonal contraceptives				<.0001			0.2702
No	N (%)	2,657 (96.5%)	1,883 (88.1%)		882 (92.7%)	869 (91.4%)	
Yes	N (%)	95 (3.5%)	254 (11.9%)		69 (7.3%)	82 (8.6%)	
Sex hormones				0.0224			0.7747
No	N (%)	2,608 (94.8%)	1,992 (93.2%)		891 (93.7%)	894 (94.0%)	
Yes	N (%)	144 (5.2%)	145 (6.8%)		60 (6.3%)	57 (6.0%)	
Antidepressant / Antipsychotic drugs				<.0001			0.5712
No	N (%)	1,999 (72.6%)	1,416 (66.3%)		697 (73.3%)	686 (72.1%)	
Yes	N (%)	753 (27.4%)	721 (33.7%)		254 (26.7%)	265 (27.9%)	
Immunosuppressants, ex glu cocorticoids				0.0003			1.0000
No	N (%)	2,737 (99.5%)	2,103 (98.4%)		944 (99.3%)	944 (99.3%)	
Yes	N (%)	15 (0.5%)	34 (1.6%)		7 (0.7%)	7 (0.7%)	
Glucocorticoids				0.0381			0.8449
No	N (%)	2,353 (85.5%)	1,781 (83.3%)		815 (85.7%)	812 (85.4%)	
Yes	N (%)	399 (14.5%)	356 (16.7%)		136 (14.3%)	139 (14.6%)	

THE PROPENSITY SCORE MODEL

The likelihood of a subject receiving treatment B was obtained using the proc logistic procedure which fits linear logistic regression models for binary or ordinal response data by the method of maximum likelihood ⁷.

The SAS® code:

```
proc logistic data=fullbase;
class &varlist/ order=freq param=ref ref=first;
model indextype(event="TRTB")= &varlist;
output out=prop_score p=prop_score;
run;
```

DESCRIPTION OF THE PROC LOGISTIC CODE

Data= the dataset to be analysed.

Class=classification variables to be used in this analysis i.e. the categorical variables.

Model= the statement used to specify the model for the analysis. The first variable **indextype** is the treatment classification of 'TRTA' or 'TRTB' and after '=' is the macro variable &varlist which contains all of the variables selected from the baseline characteristics, comorbidities and the non-epileptic drugs. The output dataset is called prop_score and the p= names the variable containing the predicted probabilities i.e. the propensity scores.

Some of the results of the logistic regression are shown in Table 3. Some of the covariates such as age at index date, the patient's gender, the index date calendar year were statistically significant and others were not such as the Germaine Smith epilepsy –specific Comorbidity index over the pre-index period. However, all of the variables were used for the matching of the subjects as the purpose of running the logistic regression was to produce the propensity scores.

Table 3 - Logistic Parameter Estimates – Propensity for being prescribed Treatment B – Model Output

Variable	Odds Ratio Description	Odds Ratio Estimate	Lower 95% Confidence Limit for Odds Ratio	Upper 95% Confidence Limit for Odds Ratio	Pr > Chi-Square
Age at index date (Years)	(x+1) / x	0.970	0.964	0.975	<.0001
Patient's Gender	Male vs Female	0.546	0.462	0.645	<.0001
Smoking status	Smoker vs Non-Smoker	0.904	0.729	1.120	0.3559
Index date calendar year (rescaled 0-9)	(x+1) / x	1.329	1.289	1.370	<.0001
Epilepsy type	Partial vs Unspecified	0.711	0.575	0.878	0.0015
Epilepsy type	Generalised vs Unspecified	0.932	0.718	1.208	0.5932
Germaine-Smith epilepsy-specific Comorbidity index over pre-index period (V1 Codes)	(x+1) / x	0.982	0.761	1.266	0.8867

MATCHING APPROACH

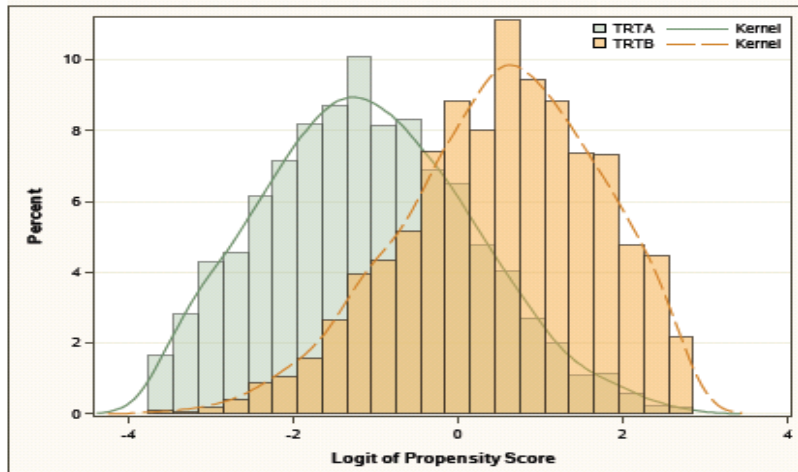
The Propensity Score (PS) was transformed to $\text{logit}(\text{PS})$ as it has been shown that a greater match can be obtained by using this transformation of the PS and matching within PS calipers is more accurate than matching on just the raw PS³. This was implemented as follows:

```
data match1;
set prop_score;
logitps = log(prop_score/(1-prop_score));
run;

proc sql noprint;
select 0.2*std(logitps) into: caliper from match1;
quit;
run;
```

Thus, the caliper was defined as 0.2 times the standard deviation of $\text{logit}(\text{PS})$ (resulting in a caliper of ~0.57). Within each age band, available patients were matched 1-1 on their $\text{logit}(\text{PS})$ +/- caliper between the two cohorts. Using the caliper method gave a good overlap between the two cohorts and more patients were matched (Figure 1).

Figure 1 Bar Chart of propensity scores by Cohort by 0.1 score range and Histogram of Logit PS



The matching of patients was performed using an in-house macro called %match. Before the %match macro could be employed to perform the matching, the cases, controls, $\text{logit}(\text{PS})$ and caliper values had to be derived:

Manipulation of the data before %match was called:

```
data cases controls;
set match1;
if indextype='Treatment B' then output cases;
else output controls;
run;
```

The purpose of %match is to create control and case datasets by matching on the specified attributes :

```
%match
(case_ds=cases,
control_ds=controls,
patient_id=patid,
match_num=1,
attributes=ageband,
attributes_range=logitps,
attributes_range_values=&caliper,
out_ds=match2);
run;
```

Description of the variables

case_ds – specifies the dataset of cases

control_ds – specifies the dataset of the controls

patient_id – specifies the variable that identifies the patient

match_num - Number of controls to match to (default 1).

attributes - Attributes to match exactly on

attributes_range - Attributes to match on a range, use in conjunction with Attributes_range_values.

Attributes_range_value - Values +/- to match range attributes on, use in conjunction with Attributes_range.

We used 1:1 matching without replacement, i.e. within each ageband 1 control was matched with one case on logit(PS) (+/- 0.2*std deviation of the logitps) in the CASES dataset to the pat_id in the CONTROL dataset. In other words, we matched on both the PS and age. There was an additional matching on age because the propensity scores were not adequate enough to account for the difference in age between the two treatment groups. We still had significant differences between the groups on the PS-matched cohorts, mainly the elderly and youngest age categories.

The treatment differences were re-calculated for all of the covariates after matching. Table 1 shows that the covariates which were statistically significant before matching no longer showed statistical significance after matching, and the number of subjects in each treatment group were similar.

CONCLUSION

This type of analysis can be a challenge for a programmer as guidance is needed on which covariates should be selected for the propensity score model and which matching approach should be employed in the study. The programmer can advise the team on which matching approach can be used based on previous experiences and performing exploratory analyses, however the final decision needs to be made by the study team. Once this decision has been made the programmer can work independently on the rest of the study.

We have shown that propensity score matching can be successfully employed in an observational study. Initially there were 4,889 patients eligible for the study. After matching there were 951 patients in each treatment group, which is a reasonable sample size for the purpose of the study.

318 variables were identified as potential source for confounding. To match on all of them separately would not have been possible, therefore applying the method of propensity score matching was the chosen approach. To match on a single value, rather than taking into account many variables with many levels simplified the matching process.

After matching all covariates were balanced across treatment groups. Hence, it can be concluded that the matching of patients was successfully carried out in this study and that the subsequent results from the cost difference between the two treatment groups were based on subjects with similar characteristics at baseline and therefore were not subject to observed confounding.

REFERENCES

1. D'Agostino, RH (1998). Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group, *Statistics in Medicine*, 17, 2265-2281
2. Rosenbaum, PR and Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55
3. Rosenbaum, PR and Rubin, DR (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score, *The American Statistician*, Vol. 39, No. 1, 3338.
4. Rosenbaum, PR (2002). *Observational Studies*. Springer, New York
5. Rubin, DB (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318-324.
6. Rubin, DB (1980). Bias Reduction Using Mahalanobis-Metric Matching, *Biometrics*, 36, 293-298.
7. SAS/STAT® User's Guide, Version 8 . Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ACKNOWLEDGEMENTS

We would like to thank Simon Borghs(UCB Pharma Ltd) for letting us use his study data in this paper and John Logan (Stats4Pharma Ltd) for his help with the preparation of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued. Contact the authors at:

Author Name: Shelley Fordred
 Company: Savvy Stats Ltd
 Email: fordreds@aol.com

Author Name: Victor Kiri
 Company: FV and JK Consulting
 Email: victor.kiri@fjkconsult.com

Author Name: Knut Mueller
 Company: UCB BIOSCIENCES
 Email: Knut.Mueller@ucb.com