APPLICATION OF PROPENSITY SCORE METHOD IN GlaxoSmithKline's TYKERB BREAST CANCER STUDY

Dr. Sourish Saha

INTRODUCTION

In some studies, investigators have no control over treatment assignments, and the differences could lead to skewed estimates of treatment effects. In order to reduce possibility of biased results, a propensity score for an individual can be used. Defined as the conditional probability of being treated given an individual's covariates, the propensity score can be used to reduce the bias in observational and non-randomized studies. The propensity scores can be used in many different fields including epidemiology, health services research, clinical trials and social sciences. There are a number of good recent examples in literature that discuss the effective usage of propensity scores to reduce treatment covariate imbalances.

Randomization of study subjects to different treatments guarantees that on average there should be no systematic differences in observed covariates. That is to say in a randomized experiment there should be very little bias in between study subjects undergoing different treatments in randomized experiments. In a non randomized observations study, however, investigators lack sufficient control over the assignment of treatment and direct comparisons of study outcomes may be misleading. This may be avoided in part by incorporating if information on measured covariates is incorporated into the design of the study being conducted by matched sampling or is made part of the treatment effect through stratification or covariance adjustment. While traditional methods of adjustment are often limited by the finite number of covariates for adjustment. Propensity scores which provide a scalar summary of the covariate information are not limited in this way.

The primary reason propensity scores are used is to reduce bias and to produce more accurate outcomes. The three most common techniques that use propensity scores are matching, stratification and regression adjustment. Each of these three techniques is a way to adjust for covariates before or during the calculation of the treatment effect. While the propensity score is calculated the same way, the calculated estimate is applied differently according to the scenario. The propensity score is the conditional probability of treatment given the observed covariates. Study subjects in treatment and control groups with equal or nearly equal propensity scores will generally also have equal or near equal distributions on their background covariates. Applying the propensity score will likely remove all of the bias in background covariates. This makes it possible to use the propensity scores to achieve accurate bias adjustments and does not require the adjustment of individual background covariates.

MATCHING

Frequently, there are studies in which the number of control patients far exceeds the number of patients treated. To control background covariates, a matching technique may be used to choose control subjects who are matched with treated subjects. It is often difficult to find study subjects who are similar enough to be matched on all important covariates even when only a few background

covariates are necessary. Propensity score matching allows an investigator to control for many background covariates simultaneously by matching on a single scalar variable.

STRATIFICATION

Another method used to control systematic differences between control and treated groups is called stratification or sub-classification. Observed background characteristics are used to group study subjects into strata. Once these strata are defined, subjects from treated and control groups are compared to each other. Stratification carries with it some of the same issues as matching as the number of covariates increases. It is possible if the number of covariates is large, some strata may only include study subjects from the treated group, and predicting the treatment effect in this stratum would become impossible. Using the propensity score as a scalar summary of all the observed background covariates, stratification on it can balance the distribution of the covariates in control and treated without affecting an exponential increase in the number of strata.

REGRESSION

The propensity score is useful as a variable in regression adjustments. To adjust the final estimate of the treatment effect, one only has to find the regression of the responses on the propensity scores in both the treated and control study groups. Using the propensity score rather than performing a regression adjustment with all of the covariates used to estimate propensity score included in the model should lead to the same conclusions. One advantage to performing the two step procedure is that one can fit a very complicated propensity score model with interactions and higher order terms first. Over-parameterization should not be a concern because the goal of this propensity score model is to derive the best estimated probability of treatment assignment. Smaller models may allow investigators to conduct more accurate diagnostic checks than when many covariates are included. Covariance adjustment should be done with caution. Rubin showed that covariance adjustment may in fact increase the expected squared bias if the covariance matrices for treated and control groups are not equal. If the variance in the control group is much larger than the treated group, consider using propensity score methods for matching or sub-classification rather than covariance adjustment.

APPLICATION OF PROPENSITY SCORE METHOD IN TYKERB BREAST CANCER STUDY

Background: EGF100151 was conducted as a randomized study in 19 countries but was not conducted in China, since most of the Chinese patients could not satisfy the key eligibility criteria. EGF109491, an open label single arm study was initiated in China after results of EGF100151 was released. The Chinese Regulatory Agency and most of the investigators from China felt that it was not appropriate to conduct a randomized study where patients could be assigned to Capecitabine alone arm (knowing that lapatinib+capecitaine arm is better).

Question: After review of the Dossier, Chinese Authorities have questioned the comparability of the populations between the global study EGF100151 and the study conducted in Chinese patients, EGF109491 and particularly asked to perform propensity score analysis to compare the population.

Analysis: Propensity score method is not appropriate when applied to small sample sizes. Because EGF109491 has only 52 patients, this could lead to severe imbalance because some propensity score subclasses may contain patients from only one study. In the situation such as this present situation, the more appropriate method is to adjust for the covariates in a regression analyses. However, we performed this analysis since it was requested by Chinese Regulatory Agency.

Details: Stratification method was used to compare the two populations.

Step 1: The baseline covariate factors from both the studies were combined. The following baseline covariates were identified: age, race (Chinese vs non-Chinese), prior capecitabine, prior taxane, prior trastuzumab, ER/PR status, ECOG performance status, stage at initial diagnosis, histology at initial diagnosis, stage at screening, number of metastatic sites at baseline, height, weight, heart rate, systolic and diastolic blood pressure.

Step 2: A group comparison was made prior to stratification using t-test (for continuous variables) or Chi-square test (for categorical variables). The test results are shown in Table 1 below. The significant factors have been marked with red.

Table 1: Group comparison prior to stratification

Variable	Chinese		Non-Chinese		Comparisons	
Total Patients	N=58		N=192			
	Mean	SD	Mean	SD	Test statistic	P - value
					t-statistic	
Age	48.41	10.69	53.90	10.42	3.49	0.0006
Stage at initial diagnosis	2.58	0.78	2.53	0.80	-0.36	0.7168
Histology at initial diagnosis	16.19	19.45	22.18	27.86	1.53	0.1280
Stage at screening	3.98	0.13	3.96	0.19	-0.73	0.4682
Height	159.25	4.75	161.12	6.67	1.96	0.0515
Weight	61.07	8.99	69.33	14.34	4.06	<0.0001
Systolic BP	122.82	13.00	126.63	15.66	1.65	0.1009
Diastolic BP	74.52	8.53	77.12	11.15	1.60	0.1106
Heart beat	81.77	8.49	81.81	12.83	0.02	0.9809
	N**	%	N**	%	Chi-square statistic	
Prior Anthracycline	50	20	188	75.2	13.36	0.0003
Prior Capecitabine	14	5.6	0	0	49.09	<0.0001
Prior Taxane	54	21.6	192	76.8	13.46	0.0002
Prior Herceptine	39	15.6	190	76	58.24	<0.0001
ER/PR Status	26	10.4	95	38	0.57	0.4488
ECOG	27	10.8	70	28	2.04	0.1535
No. of metastatic sites	33	13.2	98	39.2	0.61	0.4340

Note: There were imbalances between on 6 of 16 baseline covariate factors.

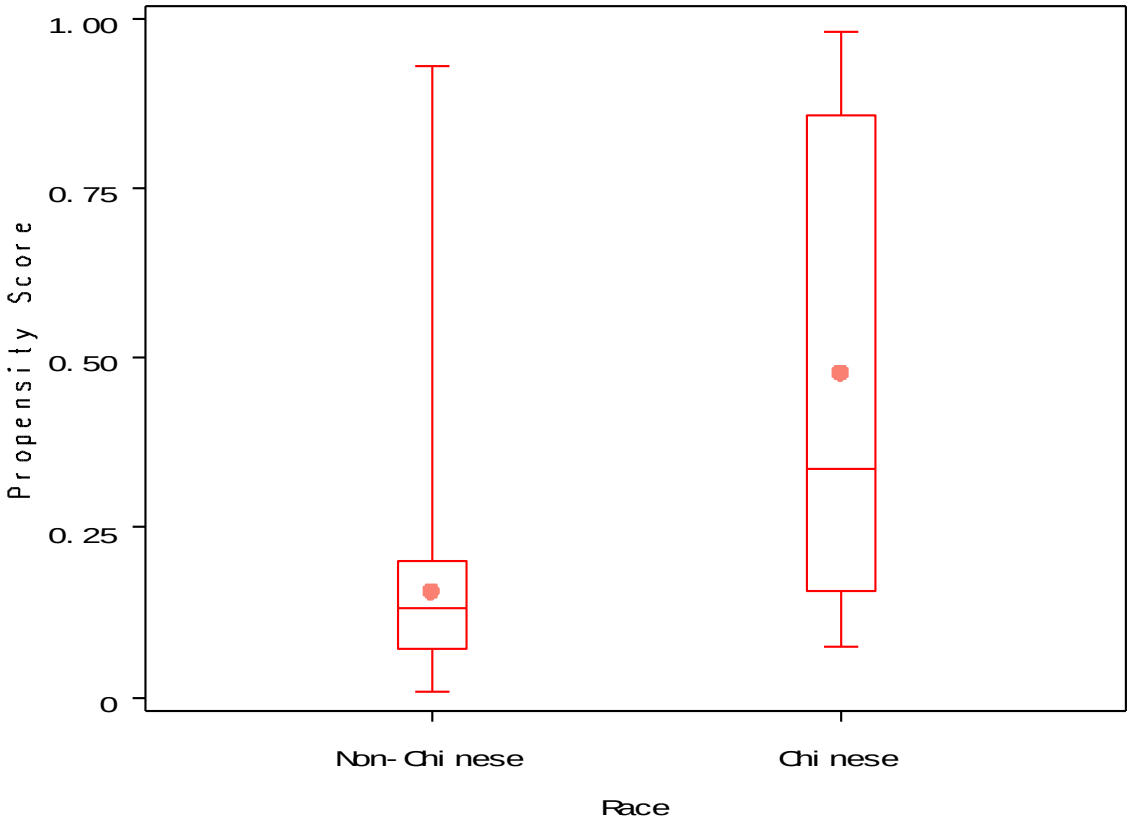
Step 3: Propensity score, the probability of being in the treated group, conditioned on the covariates, is calculated using a stepwise logistic regression is performed to estimate the propensity score for each subject having race as dependent variable. We found the following variables have imbalances between treatment arms:
Age, Weight, Prior Trastuzumab, Prior Anthracycline.

Step 4: Patients were assigned to one of 5 strata based on propensity score. Cochran (1968) had calculated that stratification based on five strata on a covariate eliminates 90% of bias in observational studies and Rosenbaum and Rubin followed his logic and argument by suggesting splitting the propensity score into quintiles in order to reduce bias. Quintiles of the estimated propensity score for the combined group were used to determine the boundaries of the strata.

Step 5: An ANOVA (Analysis of Variance) was performed to determine if the propensity score method could remove initial bias. It has been observed that after stratification on the propensity score only 2 factors were significant.

Discussion: It is not reasonable to conduct propensity score analysis since the sample size of EGF 109491 is small compared to EGF100151. This may lead to biased comparison between two groups, since some propensity subclasses may contain patients from only one group.

Figure 1: Box plots of estimated propensity score



Comment: There is very little overlap in the distribution of propensity scores between Chinese and Non-Chinese groups.

Table 2: Distribution of patients at the five propensity score quintiles

Stratification	Race	N
Quintile 1 [0.01, 0.07)	Chinese	0
	Non-Chinese	41
Quintile 2 [0.07, 0.13)	Chinese	7
	Non-Chinese	35
Quintile 3 [0.13, 0.18)	Chinese	7
	Non-Chinese	34
Quintile 4 [0.18, 0.31)	Chinese	8
	Non-Chinese	34
Quintile 5 [0.31, 0.98)	Chinese	26
	Non-Chinese	15

Comment: As we can see from the above table, the first quintile group does not contain any patient from the Chinese population, the sample size of EGF109491 being small compared to EGF100151.

SAS Codes:

```
/* Perform a stepwise logistic regression to estimate propensity scores for
each subject */;
/* The variable pr is the propensity score */;


proc logistic data=base_all nosimple;
model racecd(event='1')= age ant tax tra erpr ecog sgcadgcd diaghycd
sgcasccd metsite height weight sysbp diabp heart/selection=stepwise;
output out=preds pred=pr;
run;

/*Creates quintiles based on the estimated propensity scores */;

proc rank groups=5 out=r;
ranks rnks;
var pr;
run;

data a;
set r;
quintile=rnks+1;
```

```
run;

/* This will show the breakdown of subjects by race and propensity score
quintiles */;

proc freq data=a;
tables quintile*racecd;
run;

/* Range of Propensity Score by Quintiles */;

ods listing close;
ods output Summary=Summary_prop;
proc means data=a min max;
var pr;
class quintile;
run;
ods output close;
Ods listing;

/* Quintile means for variables */;

data no_miss;
set a;
where quintile ne .;
proc sort; by quintile;
run;

ods listing close;
proc ttest data=no_miss; class racecd;
var age ant cap tax tra erpr ecog sgcadgcd diaghycd
sgcasccd metsite height weight sysbp diabp heart;
by quintile;
ods output Statistics=Stat_quintile (keep=variable quintile class N mean
stddev);
run;
ods output close;
Ods listing;

data Stat_quintile;
set Stat_quintile;
where class not in ('Diff (1-2)');
run;


/* This will perform a 2-way novas to determine whether the propensity score
quintiles removed the initial bias found by the t-tests above */;

proc glm data=a;
class racecd;
model  age ant tax tra erpr ecog sgcadgcd diaghycd
sgcasccd metsite height weight sysbp diabp heart=quintile racecd
quintile*racecd;
run;


/* t-test within each quintile */;
```

```
data a1;
set a;
where quintile ne .;
proc sort; by quintile;
run;

ods listing close;
proc ttest data=a1; class racecd;
var age ant cap tax tra erpr ecog sgcadgcd diaghycd
sgcasccd metsite height weight sysbp diabp heart;
by quintile;
ods output TTests=T_test_1;
ods output Statistics=Stat_1 (keep=variable quintile class N mean stddev);
run;
ods output close;
Ods listing;


data t_test_1;
set t_test_1;
where method='Pooled';
run;

data stat_1;
set stat_1;
where class not in ('Diff (1-2)');
run;


proc freq data = a1;
  tables tra*racecd / chisq ;
  by quintile;
run;
```