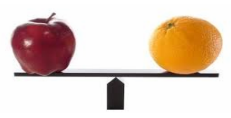
HSOR NEWS
The evidence behind your decisions
Health Services & Outcomes Research, **Issue 2011/3**

COMPARING APPLES AND ORANGES: THE USE OF PROPENSITY SCORES IN OBSERVATIONAL STUDIES

INSIDE THIS ISSUE

Two heart surgeons walk into a room. The first surgeon says: "I just finished my 100th heart surgery!" The second surgeon replies: "Oh yeah, I finished my 100th heart surgery last week. Only 10 of my patients died within 3 months of surgery." First surgeon smugly responds: "Only 5 of mine died, so I must be the better surgeon." Second surgeon says: "My patients were probably older and had a higher risk than your patients."

Such non-randomised comparisons give rise to apples-and-oranges scepticism.



Observational Studies and Selection Bias

Except by chance, patients in different comparator groups tend to differ in non-randomised studies. Although randomisation can produce relatively comparable treatment groups, observational studies are used because:

- Randomised controlled trials with strict inclusion criteria may have limited external validity
- Data are widely (increasingly) available and can reduce cost and time to get answer
- They enable examination of real life situations
- Large sizes permit investigation of exposures with smaller effect sizes

Observational studies are common in health services research but as treatment assignment is outside the control of the investigator, the potential for biases is higher.

Traditional methods to adjust for selection bias

Matching: We can match treatment and control group patients based on selected characteristics using a case control design. However, this becomes impractical when there are a large number of covariates.

Stratification: We can stratify the treatment and control groups according to selected characteristics. However, the number of sub-categories will increase exponentially, leading to a problem of having few subjects in each cell to have a meaningful comparison.

Multivariate Adjustment: We can statistically adjust for baseline differences between the groups simultaneously. However, this does not resolve the issue that the people getting treatment may be systematically different from the control group in ways that affect outcome.

What is propensity score (PS)?

Propensity score (PS) helps to balance the data sets by transforming an apples-to-oranges outcomes comparison into an apples-to apples comparison. For each patient, we can estimate the propensity toward ($0 \leq PS \leq 1$) belonging to the treatment group versus non-treatment.

To derive the propensity scores, we first construct a logistic regression model that represents the treatment allocation decision. Therefore, we should consider including any variables that have a relationship to the treatment decision. After which, we need to check whether the scores for the comparator groups overlap reasonably (Fig. 1). If they do, we can proceed to use the scores in the following ways. If not, it just means the two groups are too different to offer any sensible comparison.

Ways to apply propensity scores

We can use the PS in our analysis of treatment effects through matching, stratifying and inclusion of the score or the strata in a regression equation.

Matching: We can match the score for the first treatment patient to all control patients within a given caliper around the score. Matched patients may not be exactly similar but overall probability of being treated is close.

Stratification: PS can be stratified. The number of strata will depend on how many participants are available. Subsequent analyses can be conducted within each strata.

Regression adjustment: The continuous PS or quintile can be included as a dependent variable in the final model for testing the effectiveness of treatment.

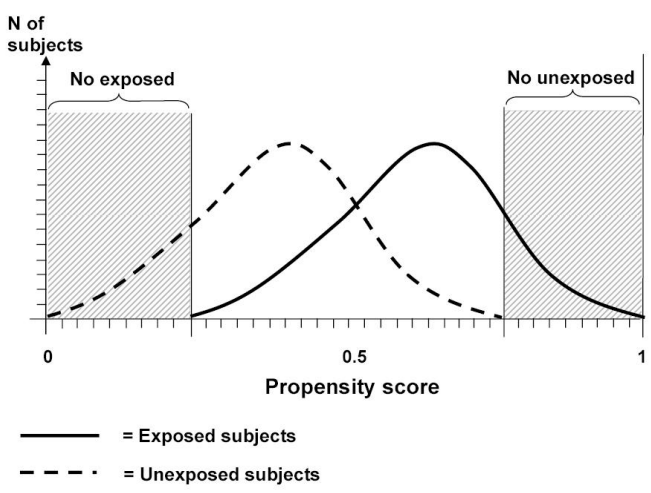


Figure 1: Non-overlap of the propensity score distributions among exposed and unexposed subjects*

** In this example subjects with low propensity scores are never exposed while subjects with high propensity scores are always exposed*

Ref: Stürmer T et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006 May;59(5):437-47

"Does aspirin use convey a survival benefit in patients with known or suspected coronary disease?"

To answer this question, a prospective, observational cohort study was conducted. Because aspirin use was not randomly assigned, potential confounding and selection biases were accounted for by developing a PS for aspirin use. A logistic regression equation with 34 covariates was used to predict the PS of aspirin use. Each aspirin user was matched to a non–user with the closest PS.

Table 1 illustrates that before PS matching, the baseline characteristics were dissimilar. The profile of aspirin users indicated higher mortality risk. However, post-PS matching, we see an apple-to-apple comparison with differences becoming statistically insignificant.

Table 1: Selected Baseline Characteristics Before & After PS Matching

	Before Matching (%)			After Matching (%)		
Variable	Aspirin (n=2310)	No Aspirin (n=3864)	P value	Aspirin (n=1351)	No Aspirin (n=1351)	P value
Male	77.0	56.1	<0.001	70.4	72.1	0.33
Diabetes	16.8	11.2	<0.001	15.0	15.3	0.83
Hypertension	53.0	40.6	<0.001	50.3	51.7	0.46
Prior coronary artery disease	69.7	20.1	<0.001	48.3	48.8	0.79
Congestive heart failure	5.5	4.6	0.12	5.8	6.6	0.43
Beta-blocker usage	35.1	14.2	<0.001	26.1	26.5	0.79
Ace Inhibitor usage	13.0	11.4	<0.001	15.5	15.8	0.79

In a simple unadjusted comparison, there was no association between aspirin use and mortality (4.5% versus 4.5%). However, in the multivariate analysis, aspirin use was associated with reduced mortality. In further analysis using matching by propensity score, 1351 patients who were taking aspirin were found to be at lower risk for death than 1351 patients not using aspirin (4% versus 8%).

Ref: *Gum PA et al. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease. JAMA. 2001 Sep 12;286(10):1187-94.*

Final word

PS cannot replace properly designed, ethical randomised trials. However, when randomisation is not practical, by explicitly modelling the treatment allocation process, it helps the analyst ensure an apple to apple comparison.

Limitations

However, relatively large sample sizes are required to facilitate the use of PS in stratified or matched case analyses. Furthermore, when important variables influencing selection are not collected, the PS may not reflect the selection process, and will be seriously degraded. Hence, to adjust for unobserved covariates, we may have to look to other econometric methods such as Instrumental variables (IV) or Differences-in-Differences (DID).

Woan Shin, Principal Research Analyst, has been a principal investigator on several projects. She has conducted technical analyses for studies evaluating quality improvement interventions, chronic disease interventions, health resource utilization and costs. Her current research interests include evaluating the impact of risk adjustment methods on outcomes and the application of economic evaluations in healthcare. She was one of three recipients of the ISPOR Contributed Research Award for Best New Investigator Podium Presentation in 2007. Prior to joining NHG, Woan Shin worked as a Research Economist with the Ministry of National Development

Tan Wan Shin, BSocSc (Hons) (Economics), MSocSc (Economics)

Feedback and enquiries: hsor@nhg.com.sg

8 June 2011