Propensity Scores: mixing & matching

Jonathan Alsop

Numerus Ltd

ROeS, 10 Sep 2013

Background

- Propensity scores - method to estimate treatment effect when random assignment not feasible / possible (e.g. observational studies)
- Subjects with certain characteristics more likely to receive treatment than others.

Are differences in outcome between treated & control subjects due to treatment OR differences on other characteristics?

Context

Do drugs/devices work in 'real-life' settings?

FDA Sentinel Initiative
- 2007
- AE reporting from *100 million* patients

EMA PROTECT
- 2009
- monitoring benefit-risk of medicines

CPRD (NHS, UK)

Effectiveness, but safety signals: rare & drug-drug

Nutshell

Compare treatment between groups who "looked similar" prior to treatment assignment

Groups selected/analysed based on PS:

conditional probability a subject would receive a certain treatment based on his/her pre-treatment characteristics

treated subject PS = control subject PS

⇒ observed covariates are controlled for

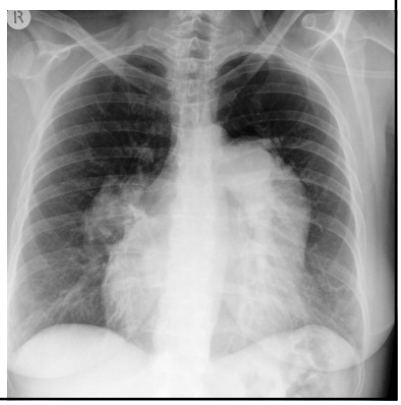
Create pseudo-randomised study

So who's to blame?

Rosenbaum and Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Casual Effects."

Case study: PAH Registry

- Pulmonary arterial hypertension
- Progressive disease - narrowing of arteries
- Prevalence of 15-50 / million
- 20 - 50% dead < 3 years



PAH Registry

Is combination therapy better than monotherapy?

Outcome = time to death from diagnosis

Point of diagnosis (demography, disease severity)

→ Initial monotherapy treatment

→ All patients assessed @ 3 months

↗ Responders stay on monotherapy?

↘ Non-responders switched to combo?

Step 1: Obtaining PS

Obtain PS for each subject from:

- Logistic regression
- Classification and Regression Tree Analysis
- Neural networks

Can be done blindedly (i.e. before outcome examined)

Obtaining PS

Using logistic regression to obtain PS:

- Kitchen sink approach: can have complicated model
- Consider including interactions and polynomial effects
- 100s of covariates considered (see Hd PS)

PS 'Philosophies'

- Expert-judged treatment drivers
- Covariates that:
  - just predict the treatment assignment
  - potentially related to the outcome
  - associated with both the treatment assignment & outcome

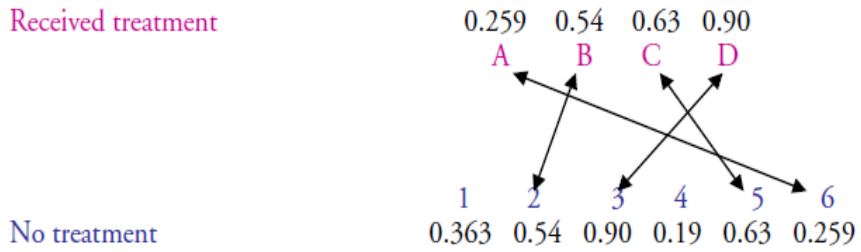
PAH Study: Obtaining PS

- N=173 patients with “Month 3” assessment
- Logistic regression (PROC GENMOD)
- Outcome: therapy type (mono=0, combo=1)
- Covariates:
  - Sex
  - Functional class
  - First line therapy (diagnosis)
  - Total Pulmonary Resistance (TPR) at Month 3
- Obtain predicted probability of combo (PS)

Step 2: Can we match?

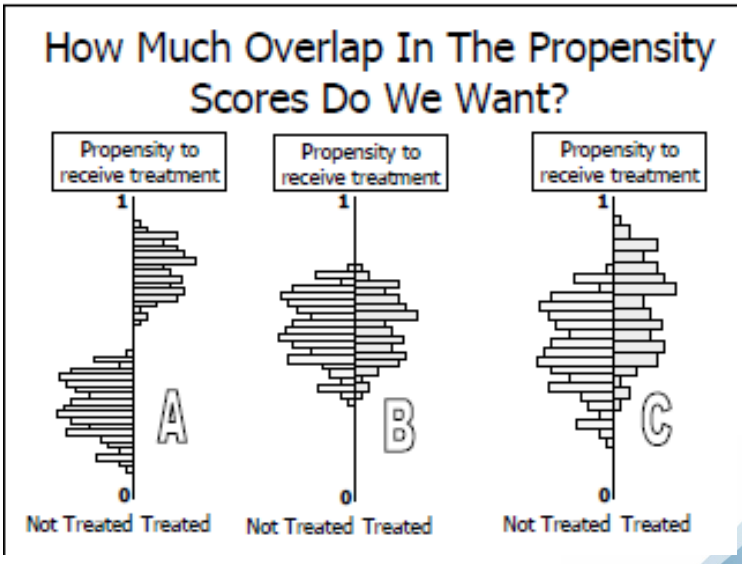
Simple matching example

Match treated subjects PS with untreated subjects with same/similar PS, e.g. 1:1 match



The unmatched subjects are discarded from the analysis

PS Overlap



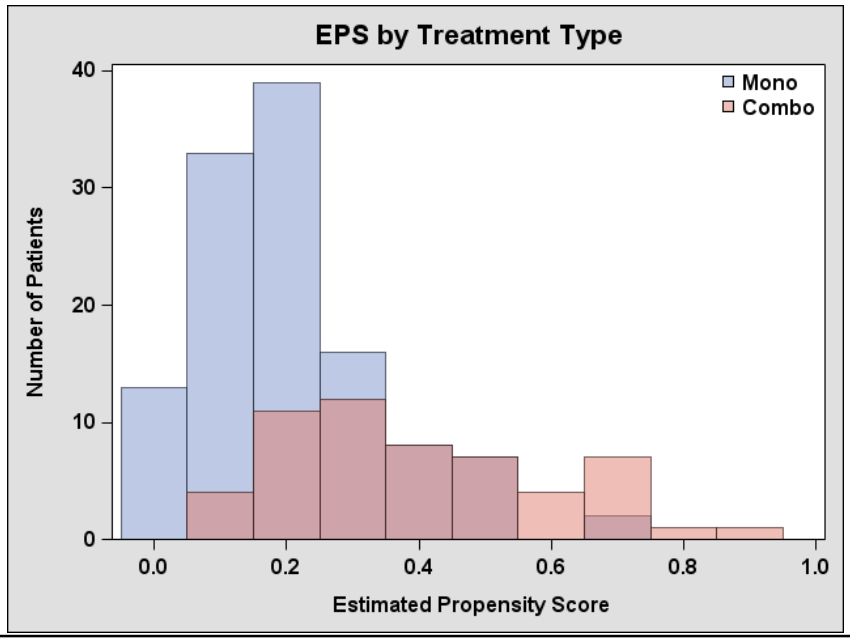
PS Overlap

Need 'substantial' overlap between treated & control otherwise…..significant loss of data / bias

Often better to match all treated patients, then follow with analytical adjustments for residual imbalances in the covariates

but see Baser 2007 – less weight on extreme PS

PAH Study: PS Overlap



Step 3: How to match/analyse?

numerus

What we really are aiming for…

numerus

Methods

One *or combination* of the 4 main methods:

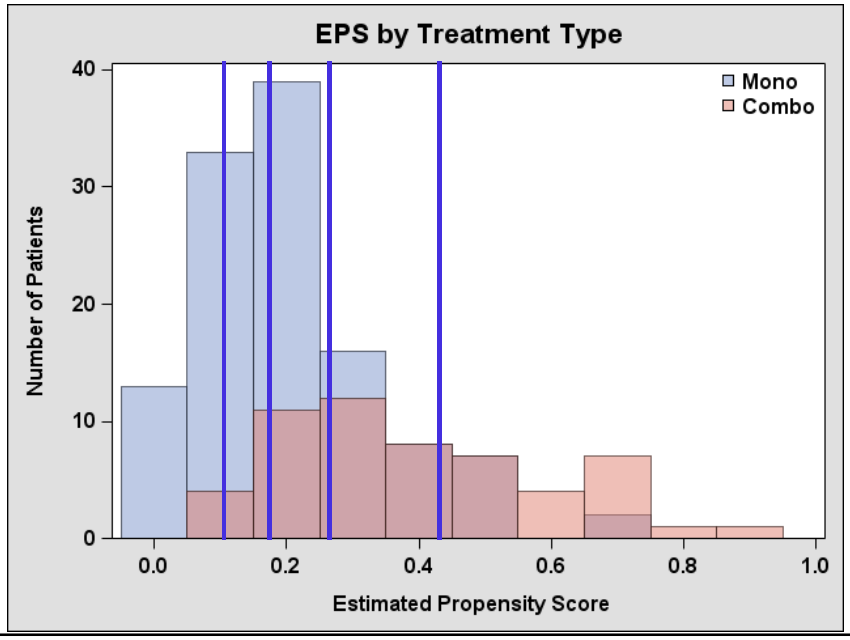
- Stratification
- Matching
- Regression adjustment
- Weighting (IPWT)

Methods: Stratification

- PS used to stratify the subjects into homogenous subclasses (e.g. deciles, quintiles) with similar PS.
- Each stratum consists of similar # of subjects *(treated? controls?)*
- Check estimated treatment effect size in each strata

e.g. PROC XXX with strata command

PAH Study: Five Strata



Methods: Matching

Treatment & control patients are matched on their PS

- Ratio treated:control (1:1, 1:n)
- With/without replacement (with has impact on analysis method)
- Caliper width (match vs. non-match)

  PS diff.=0.00001 to 0.1 ; 0.2 × SD [logit(PS)] ; 0.6 × SD(PS)
- Matching method:

  Nearest neighbour (with replacement)

  Greedy (nearest neighbour without replacement, sub-optimal)

  Radius (nearest neighbours within caliper)

  Overall distribution (*network flow theory*, optimal, PROC NETFLOW SAS/OR)

  Kernel (local linear), Genetic,...etc!

SAS progs: Parsons (2004), Fraeman (2010),
http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm

Adjust analyses for matches! ***e.g. paired t-test, stratified, GEE/multi-level***

Matching: Checking covariate balance

- Propensity score dist.
- Standardized differences
- Compare covariate distributions (e.g. boxplots, q-q plots)
- Regression (e.g. ANOVA on covariate)

Methods: Regression & IPWT

Covariate/Regression adjustment

- PS is a covariate (cont. or cat.) in regression model

Weighting (IPTW aka Inverse Probability of Treatment Weighted)

- patients are re-weighted in order to make them more representative of the population *e.g. PROC XXXX with weight command*
- weight: Treated= 1/PS, Controls= 1/(1-PS)

  e.g. treated subject with a 20% predicted probability =
  4 × treated subjects with a 80% predicted probability
- sensitive to outliers - truncation/trimming (e.g. at 5-95 pct)
- variance(tx effect) ideally take into account var(PS)
  - bootstrap from PS model onwards

Ok, so which method?

Combination of methods often used

No method is deemed 'best'

I vote for…

1st Greedy matching with covariate adjustment

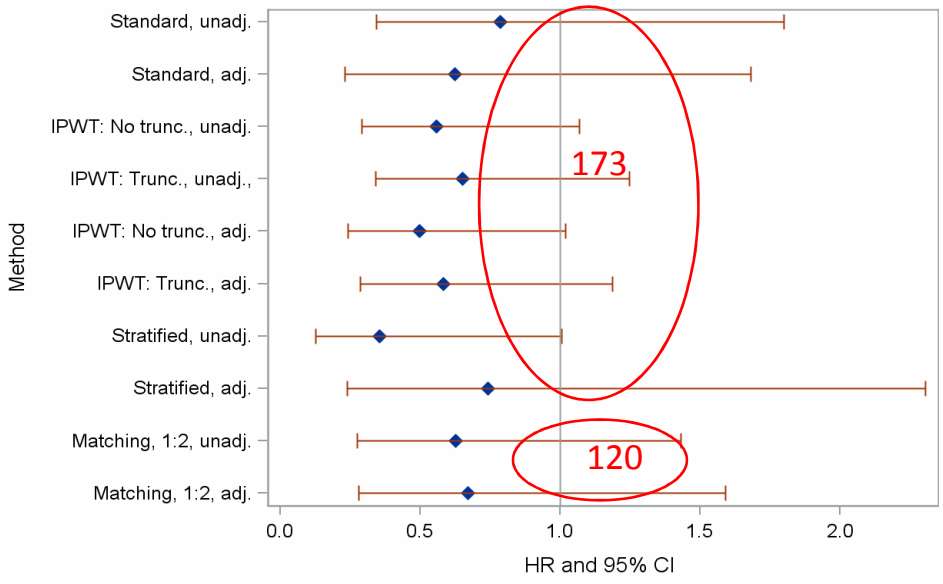
2nd Stratification with covariate adjustment

PAH Study: Matching PS

- *Standard Cox PH, time to death* PROC PHREG on all pts
- Stratified 5 strata, PROC PHREG on all pts; *strata* option
- Matching PROC PHREG on sample matched pts ; *strata* option
  - Greedy, 1:2 combo:mono, max diff=0.1
- IPWT
  - PS weights PROC PHREG on all pts; *weight* option
  - with/without truncation truncated at 5.7 (90 pct)

All: with/without covariates

PAH Study: Cox PH Results



Why Not Do 'Plain-Vanilla' Regression Instead?

Complicated PS model with interactions & polynomial terms

Simpler final model when PS is included

Rare events: PS analysis better than standard regression

Easier to check covariate balance

Extensions

Apply propensity score in longitudinal studies

Construct time-dependent propensity score:

- sequential matching
- inverse-probability-of treatment weighted (IPTW) estimator
- marginal structural models

High-Dimensional Propensity Score

- hundreds of possible treatment predictors

The good...

- Intuitive (face value, but….?)
- Predicting treatment allocation interesting
- Analyse rare events
- PSM reduces bias of treatment comparisons
- More robust than regression
- Can check balancing assumptions easily
- Can check the overlap of multivariate covariate distributions in the control and treated groups

The not so good...

- Larger variance cf regression, data loss
- PS balance achieved?
- Assumes treatment constant within a patient but see MSM
- Biased if unobserved confounders exist (unlike rand^{n})
- Large sample required
- Pts within matched pairs aren't independent see Austin (2007)
- Appropriate variance estimation and CI calculation for PS analyses often omitted (bootstrapping required?)
- Handling of missing data not clear
- Use in generalised & non-linear models less well understood
- Applicability & representation cf. population

numerus

PS analysis summary…

Useful *additional* tool, but *easy*?

Obtaining PS 

Matching / analysis 

Trade-off precision vs. bias (but surely bias more important…?!!)

Future use:

Pharmacovigilance

Blinded combination of health databases (Rassen 2010)

Phase 3b/4 trials more easily undertaken / analysed?

Stratified medicine need large, multi-tx trials (observational)

numerus

Recommended reading

A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Austin P. Statist. Med. 2008; 27:2037–2049

An Introduction to Implementing Propensity Score Matching With SAS®. Fraeman, K. 2010. NESUG

Analysis of Observational Health Care Data Using SAS. Faries D, Leon A, Haro H, & Obenchain R. 2010

High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. Epidemiology. 2009; 20(4):512-22

Performing a 1:N Case-Control Match on Propensity Score. Parsons L. 2004. SUGI-29

Propensity Score Matching with Limited Overlap. Baser O. Economics Bulletin. 2007; 9(8):1-8

Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques. Parsons L. 2001. SUGI-26

Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from multiple health care utilization databases. Rassen JA, Avorn MD and Schneeweiss S. Pharmacoepidemiology and Drug Safety 2010; 19: 848–857

The central role of the propensity score in observational studies for causal effects. R. Rosenbaum P, Rubin D. Biometrika. 1983; 70 (1): 41-55

numerus
Discover Statistical Power

Thank you!

numerus
Discover Statistical Power

Good PSM

- Matching based on variables that are accurately & reliably measured.
- Substantial overlap between groups on the PS.
- Adequately balance on covariates between groups.
- Adjusts for selection bias & minimizes group differences across many variables.
- It does not use only conveniently available covariates such as age and gender.
- Sensitivity analysis is a recommended part of the process
- Choosing variables and adjusting for propensity scores is based on logic, theory & empirical evidence

Comparing PSM with Hard Matching

- PSM is more suitable when dealing with a large number of covariates whereas hard matching is more appropriate when dealing with a small number of covariates.
- Both methods control for observed covariates and do not account for bias resulting from the unobserved covariates that may affect whether a subject receives treatment or not.
- PSM and matching both produce similar results when matching on a small number of covariates.

PS Overlap

What if Treated and Untreated groups overlap, but minimally?

- Not much help
- The info available to infer treatment effect will reside almost entirely in the few patients who overlap.
- Need to think hard about whether useful inferences will be possible.

Step 4: Check Sensitivity

"How much hidden bias would have to be present to alter the study's conclusions?"

Table 5: Representative results of sensitivity analysis* on time from first visit to end-points SP and EDSS 4.0 and 6.0: how the magnitude of an unmeasured binary confounder might affect the propensity score-adjusted HRs of Table 4

			Adjusted	
End-point	HR†	P_0-P_1‡	HR	95% CI
SP	2	0.8	0.66	0.41–1.00
	4	0.4	0.67	0.42–1.02
	6	0.3	0.67	0.42–1.02
	8	0.2	0.69	0.44–1.06
EDSS 4.0	2	0.1	0.76	0.58–1.03
EDSS 6.0	2	0.1	0.65	0.41–1.03

*This analysis assumes that: 1) the unmeasured confounder is binary; 2) the unmeasured confounder is independent of measured confounders; 3) there is no interaction between the unmeasured confounder and exposure; †Hypothetical HR of the unmeasured confounder on time to end-points; ‡Differences in prevalence of the unmeasured confounder between IFNB-treated and controls. From Trojano *et al.*[26]

Checking covariate balance