Advances in health economic evaluation methods in the absence of random allocation

LSE Health and Social Care

29.9.2016

Richard Grieve

richard.grieve@lshtm.ac.uk

Acknowledgements

- Jas Sekhon (UC Berkeley), Noemi Kreif and Stephen O'Neill (LSHTM), Luke Keele (Georgetown University, Washington), Sam Pimental (University of Pennsylvania), Steve Harris (UCL).
- Funding: Senior Research Fellowship, National Institute for Health Research
- Team for Health Economics, Policy & Technology Assessment (THETA)
- Slides available:

http://theta.lshtm.ac.uk/

content

- Context, and current state of play
- Recent developments
  - Genetic matching
- New developments
  - Near far matching
- Discussion

Context: International initiatives on observational data

- Cancer Drugs Fund (Grieve et al, 2016)

- Accelerated Access Review

https://www.gov.uk/government/publications/accelerated-access-pathways-for-medical-technologies

- EMA, Adaptive Pathways. Pilot project on adaptive licensing.

http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/03/WC500163409.pdf

- Surveillance, Epidemiology and End Results (SEER)-Medicare–linked data. Innovative Medicines Initiative (IMI) GetReal Project

https:/www.imi-getreal.eu/

- Improvements in data linkage e.g. CPRD/HES/social care/disease registries

Examples of HTA that use observational data

Evaluation	Analytical method	Reference
PTCA vs. CABG for Angina	Regression	Griffin et al (2007)
Surgery bladder cancer	Propensity score (Pscore) methods	Mitra and Indurkhya (2005)
ECMO for patients with H1N1 Alternative types of hip prosthesis	Pscore matching, Genetic Matching	Noah et al (2011) Pennington et al (2013)
Surgery for breast cancer	IV methods	Polsky and Basu (2006)
Treatments for psoriasis	Matching-adjusted indirect comparisons	Signorovitch et al (2010)
Bosutinib for Chronic Myeloid Leukaemia	Naïve comparison	NICE 2013, TA 299 NICE 2015, TA 413

Non-randomised studies: **key issue**

- Design different forms of study
  - Before and after study
  - Historical cohort study
  - Concurrent cohort study
  - Longitudinal study
- Common point: treatment allocation is non-random
- If variables explain treatment receipt and outcomes
  - Concern selection bias from confounding (endogeneity)
  - Develop design and analysis strategies to minimise confounding

Key Concern: confounding

HRT triples the risk of breast cancer, biggest ever study shows

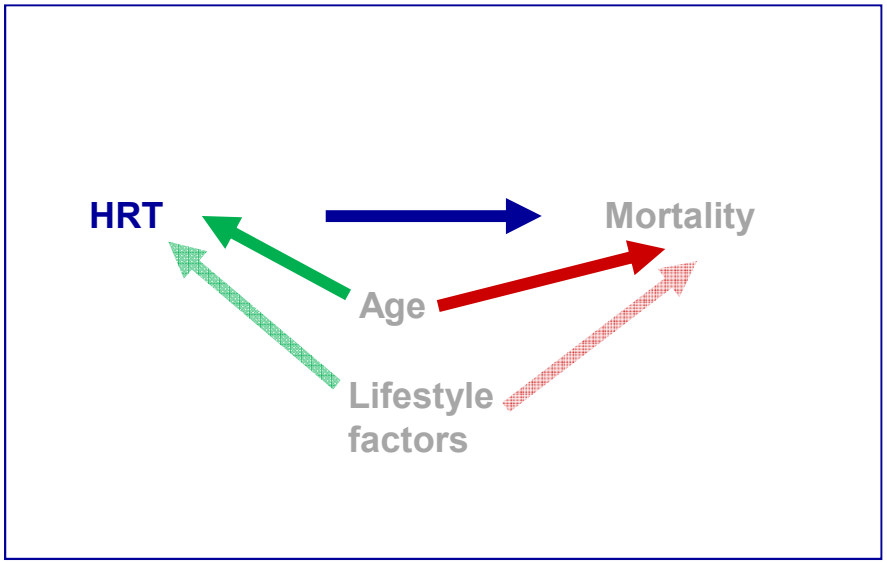
The Telegraph

'Hormone replacement therapy can triple the risk of breast cancer, the biggest ever study has found, following more than a decade of controversy. Last year NICE changed guidance to encourage more doctors to prescribe HRT claiming too many menopausal women had been left suffering in silence. Doctors were reluctant to prescribe it after a study in 2002 suggested it could raise the risk of cancer, a claim later widely disputed.
Now new findings suggest the original risk had actually been underestimated.'

A study of 100,000 women over 40 years found those who took the combined oestrogen and progestogen pill for around five years were 2.7 times more likely to develop cancer compared to women who took nothing, or only the oestrogen pill.

August, 2016

Selection bias from measured and unmeasured confounders



If ignored can lead to selection bias

- Bias from imbalance on unobservables: hidden bias

- Bias from imbalance on observables: overt bias

2. Statistical Methods for addressing confounding

- **Assume no unobserved confounding**

  e.g. Regression adjustment, matching

- **Allow for observed and unobserved confounding:**

  e.g. Instrumental variable estimation, Regression discontinuity design

How good are we at addressing confounding in health economic evaluations?

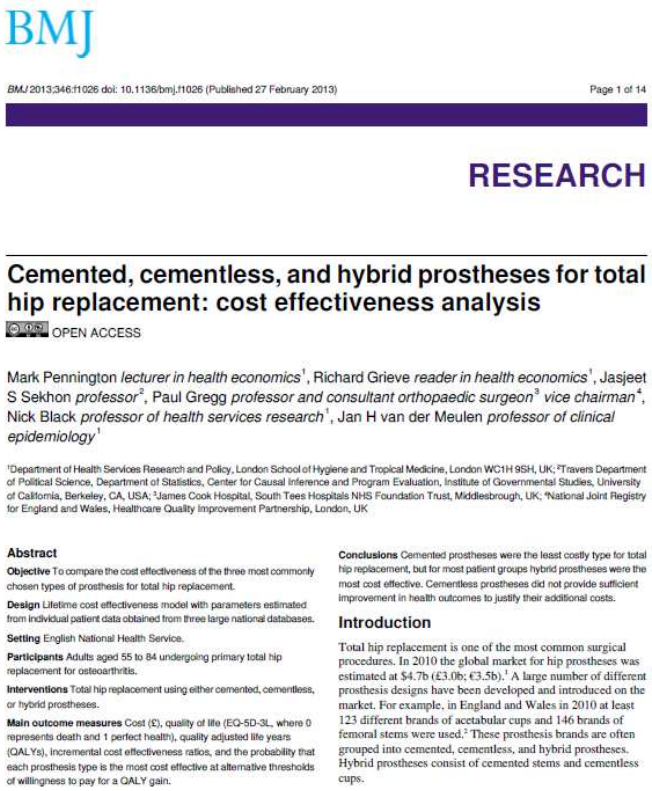
- Systematic review of economic evaluations (2000-2011) by Kreif et al (2012)
  - 79 studies
  - Almost all studies assumed “no unobserved confounding”, failed to justify
  - A couple used IV methods, did not justify identification assumptions
- Faria et al. (2015) NICE DSU
  - Single technology assessments (STAs) use matching or regression when using IPD
  - Analysis and reporting not satisfactory
  - Further training in relevant statistical methods essential

Improving the situation

Grimm et al, 2016 NICE DSU document

Careful design

- Minimises risk of confounding
- Prospective, measurement of all relevant covariates
- Exploit large, linked datasets
- Administrative data (HES,PROMs, National Joint Registry)
- Careful analysis and interpretation key too..



BMJ

BMJ 2013;346:f1026 doi: 10.1136/bmj.f1026 (Published 27 February 2013)

RESEARCH

Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis

OPEN ACCESS

Mark Pennington *lecturer in health economics*[1], Richard Grieve *reader in health economics*[1], Jasjeet S Sekhon *professor*[2], Paul Gregg *professor and consultant orthopaedic surgeon*[3] *vice chairman*[4], Nick Black *professor of health services research*[1], Jan H van der Meulen *professor of clinical epidemiology*[1]

[1]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London WC1H 9SH, UK; [2]Travers Department of Political Science, Department of Statistics, Center for Causal Inference and Program Evaluation, Institute of Governmental Studies, University of California, Berkeley, CA, USA; [3]James Cook Hospital, South Tees Hospitals NHS Foundation Trust, Middlesbrough, UK; [4]National Joint Registry for England and Wales, Healthcare Quality Improvement Partnership, London, UK

Abstract

Objective To compare the cost effectiveness of the three most commonly chosen types of prosthesis for total hip replacement.

Design Lifetime cost effectiveness model with parameters estimated from individual patient data obtained from three large national databases.

Setting English National Health Service.

Participants Adults aged 55 to 84 undergoing primary total hip replacement for osteoarthritis.

Interventions Total hip replacement using either cemented, cementless, or hybrid prostheses.

Main outcome measures Cost (£), quality of life (EQ-5D-3L, where 0 represents death and 1 perfect health), quality adjusted life years (QALYs), incremental cost effectiveness ratios, and the probability that each prosthesis type is the most cost effective at alternative thresholds of willingness to pay for a QALY gain.

Conclusions Cemented prostheses were the least costly type for total hip replacement, but for most patient groups hybrid prostheses were the most cost effective. Cementless prostheses did not provide sufficient improvement in health outcomes to justify their additional costs.

Introduction

Total hip replacement is one of the most common surgical procedures. In 2010 the global market for hip prostheses was estimated at $4.7b (£3.0b; €3.5b).[1] A large number of different prosthesis designs have been developed and introduced on the market. For example, in England and Wales in 2010 at least 123 different brands of acetabular cups and 146 brands of femoral stems were used.[2] These prosthesis brands are often grouped into cemented, cementless, and hybrid prostheses. Hybrid prostheses consist of cemented stems and cementless cups.

Recent developments in evaluation literature

I: Matching (sekhon and Grieve 2011)

- Regression correct functional form never known
- Incorrect relationship: parameter to endpoint
- Or treatment effect multiplicative not additive
- Biased and inconsistent estimates
- Especially severe when weak overlap (Ho et al. 2007)
- Regression involves extrapolation
- Endpoint variable always in sight
- Interest in matching as part of study design, i.e prior to outcome analysis

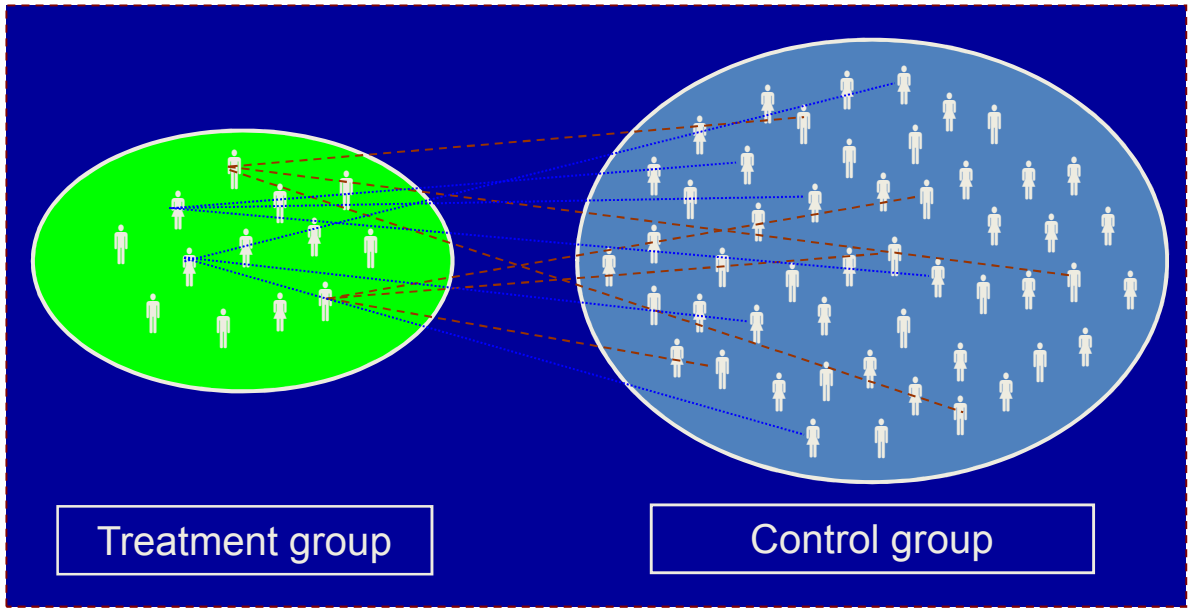
Matching (see Stuart 2010 for excellent review)

- AIM: ensure groups are balanced
- Part of general principle about importance of good design
- Distributions of baseline covariates similar between treatment and control groups
  - Means, but also variances etc
  - e.g. in RCT similar baseline covariate distributions

- Imputes missing potential outcome by finding a "similar" individual from the control group
- Similarity based on *observed* characteristics

- Key assumptions
  - 1. No unobserved confounders
  - 2. Covariates overlap between groups $0<Pr(T_i=1|x_i)<1$

Intuition behind matching
e.g. for average treatment effect for treated (ATT)



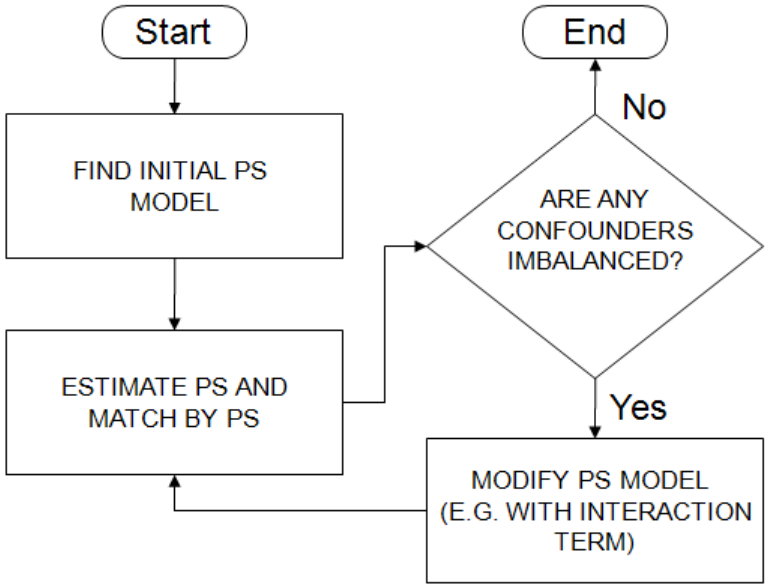
Require matching method that achieves **best balance** in observed characteristics x_i between treatment and control groups

Propensity score (Pscore)

$$e(X_i) = \Pr(T_i = 1 \mid X_i)$$

- Model of the probability of treatment, given observed covariates
- Choice of treatment depends on patient, clinician choice
- Matching on Pscore can give unbiased estimate ATT (Rosenbaum, Rubin 1983)
- If Pscore is correctly specified
  - Pscore generally unknown, must be estimated
  - Balance can be directly assessed, shows if Pscore is specified correctly
  - How do we get correct functional form?
  - Assess balance post matching, modify accordingly
  - Achieving balance on many terms is challenging..

Iterative process for specifying the Pscore



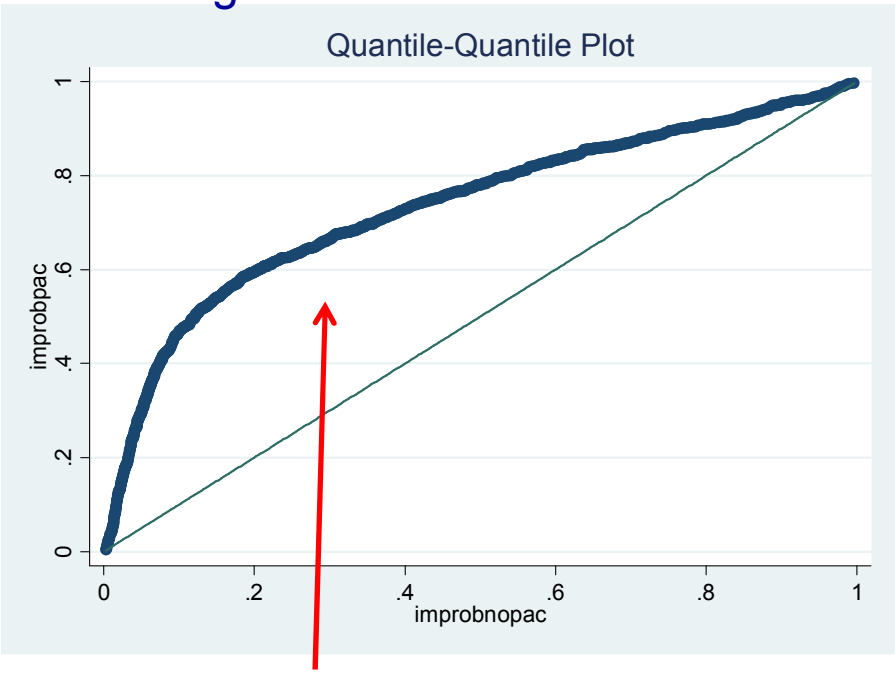
Importance of fully assessing covariate balance

Empirical Quantile-Quantile Plot (eQQ)

Pulmonary artery catheter (PAC) versus no PAC

Baseline probability death (IMProb)

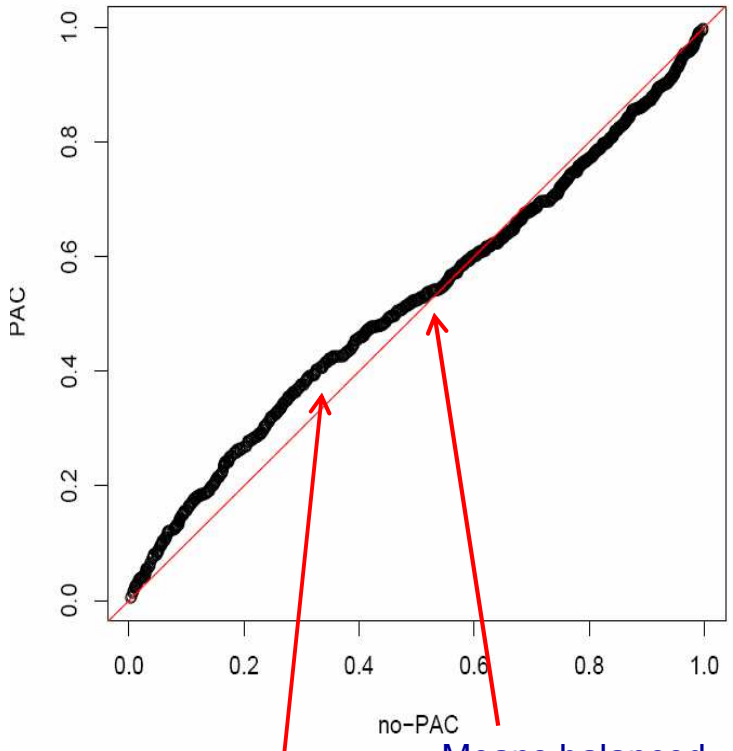
before Pscore matching



Want the gap to be small

i.e linked p value to be large

after Pscore matching



Means balanced

Still a gap, albeit smaller

Motivation for Genetic Matching (GenMatch)
see Sekhon (2011)

If balance is bad, what next?

- Modify Pscore
- Re-check balance
- When is balance good enough??
- No consensus
  - Standardised difference <10 (Austin 2009)
  - Maximise without limit (Ho et al. 2007, Sekhon 2011)

- **Aim:** max balance between treatment and controls
- **Genetic Matching**: Automated search algorithm maximises balance
- Recommended by Pscore developers (Rosenbaum and Rubin 1985)

GenMatch: Multivariate matching

(see Sekhon 2011, Sekhon and Grieve, 2011, Noah et al, 2011, Pennington et al, 2013, Sadique et al, 2011, Kreif et al, 2012; Radice et al, 2012; Ramsahai et al, 2011)

- GenMatch generalises Mahalanobis distance measure
- $GMD(X_i,X_j) = \{ (X_i - X_j)' (S^{-1/2})' \boldsymbol{W} S^{-1/2}(X_i - X_j) \}^{1/2}$
  - X_i and X_j vector of covariates for 2 different observations;
  - S is sample covariance matrix of X
  - $\boldsymbol{W}$ is a weight matrix
- Considers many alternative sets of weights
- A genetic algorithm searches data to pick the weights W
- Picks those weights that maximise overall covariate balance
- **Creates matched dataset using optimal weights**

Illustration GenMatch in CEA

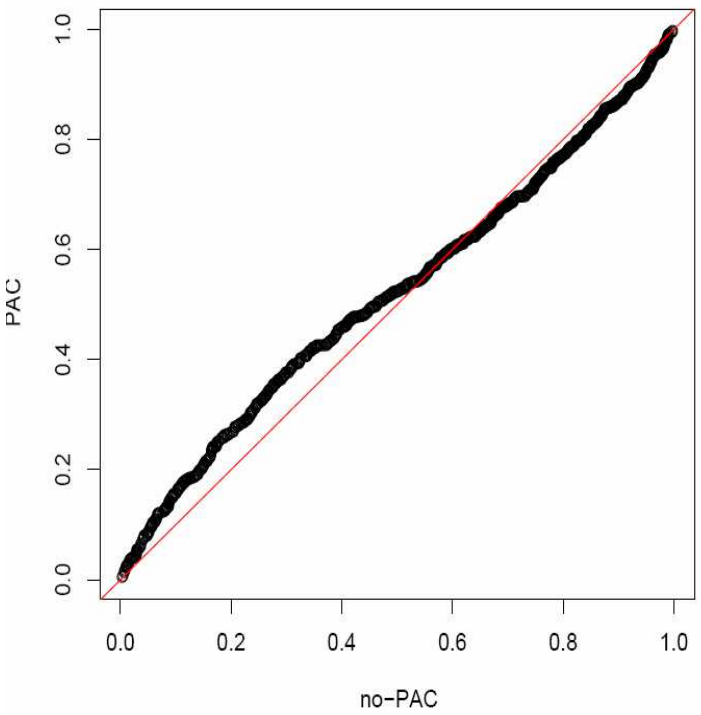
See Sekhon and Grieve 2011

- Pulmonary artery catheterization (PAC)
- Observational study using Pscore model
- PAC higher mortality & cost vs. No PAC (Connors 1996)
- Big impact on health policy: reduced PAC use
- We apply Pscore vs. GenMatch
- Also compare to estimates from RCT based analyses
- Use critical care data from ICNARC (1052 PACs, 32,000 no PACs)

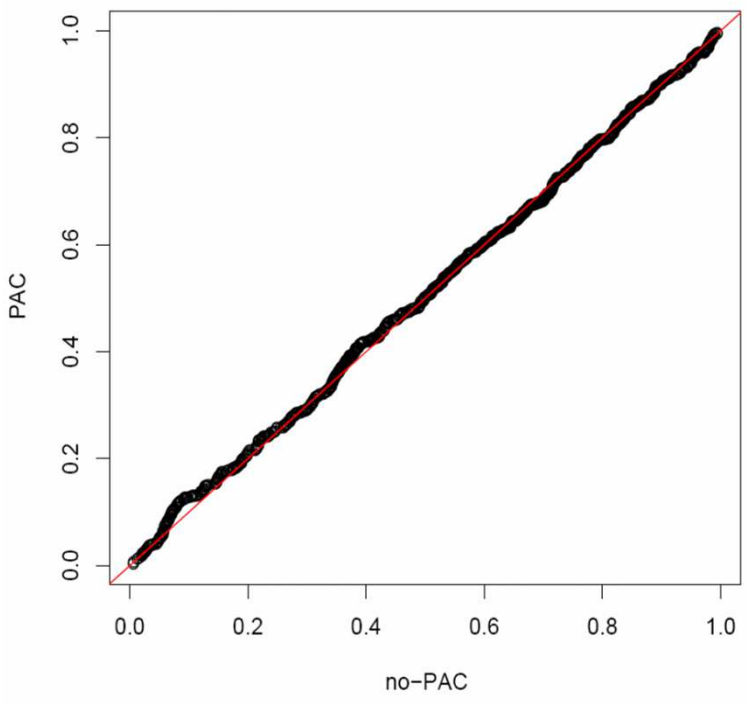
Covariate Balance: eQQ-plot

Baseline Probability Death PAC vs. No PAC

Pscore matching



Genetic Matching



Incremental net benefit (INB)
PAC vs. No PAC

	INB (95% CI)
Pscore matching	-£27,215 (-£38,864 to -£14,154)
GenMatch	-£11,830 (-£24,960 to £834)
RCT	-£3,089 (-£19,234 to £13,265)

λ=£30,000 per QALY

CIs calculated with non-parametric bootstrap

Matching alone, and combined with other regression methods

- Balance is key
- GenMatch less bias estimation ATT versus Pscore matching and IPW
  (Sekhon and Grieve, 2011; Kreif et al, 2012; Radice et al, 2012)
- Cost computational time
- Investment worthwhile with data with irregular distributions, as in CEA
- Not limited to incremental costs, or QALYs (see Noah et al, 2012)
- Further advantages combining matching with regression (Kreif et al. 2013), e.g. extends to survival analysis
- Matching combined with regression performs at least as well as double robust estimation (Krief et al 2014)
- Throughout, overarching design cross sectional data
- Assumed no unobserved confounding..

Recent Developments II: Longitudinal data

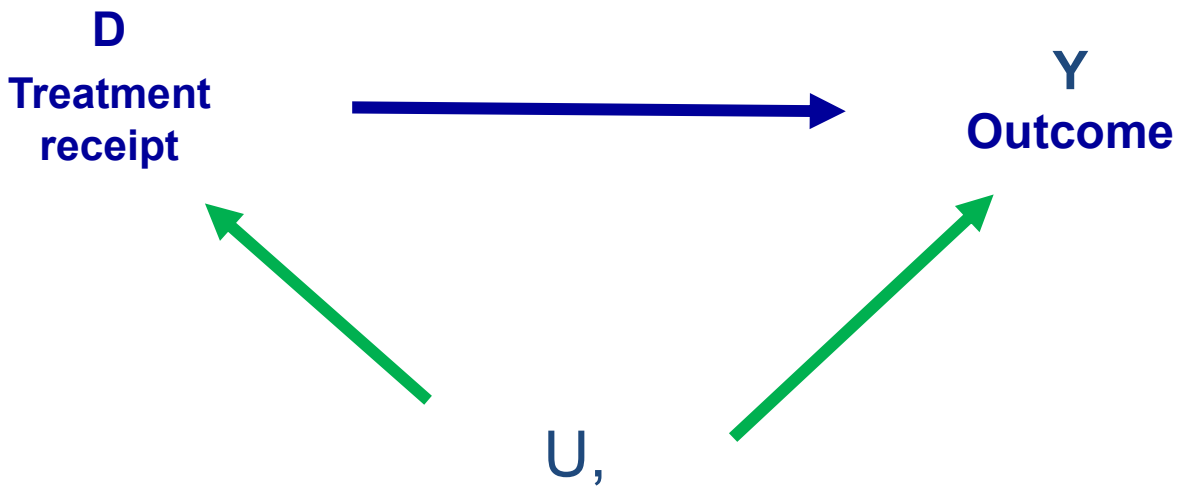
(see O'Neill et al, 2015)

- Evaluation of health policies with longitudinal data
- Could apply CEA of health policies/area level interventions (Meacock et al, 2013)
- Traditional DiD estimation, but identification assumes 'parallel trends'
- Synthetic control method, identification allows aspects unobserved confounding
- Evaluation of P4P reported different results (Kreif et al, 2014)
- Simulation, DiD vs synthetic controls vs lagged dependent variable (LDV) approach
- LDV method, recommended as alternative to DiD estimation
- Synthetic control method could be inefficient
- Design lesson: few pre-treatment periods, all methods were biased

New developments: IV methods

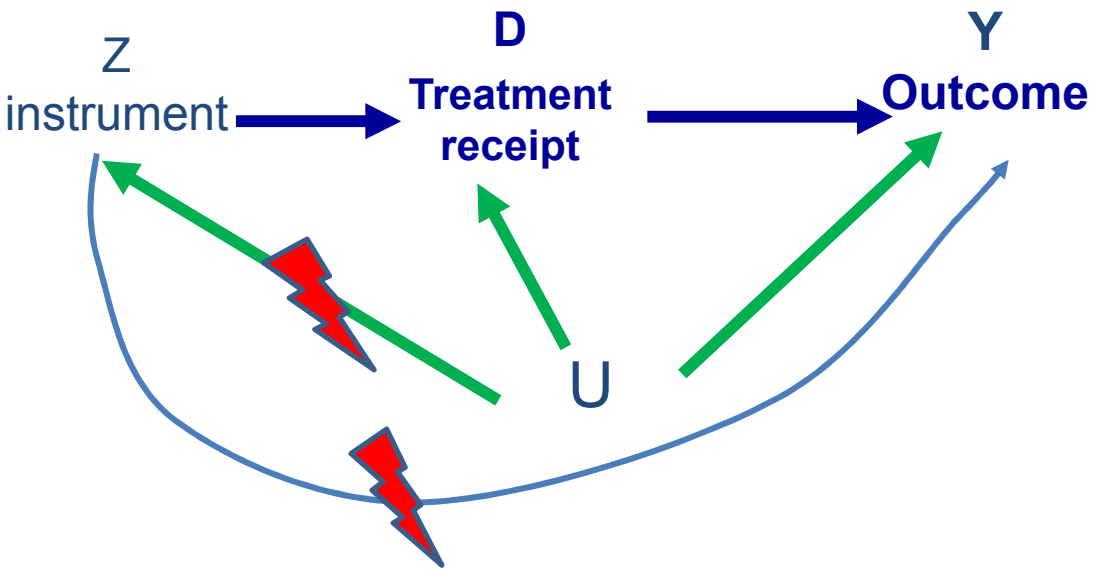
- Purport to allow for unobserved as well as observed confounding
- Natural experiments
- Potential IVs distance, prescribing choices, supply side constraints
- Progress for estimating heterogeneous causal effects (Basu et al, 2013)
- Common concern is instrument may be ‘weak’
- Weak IV consistent estimates, truly independent unobserved confounders (U)
- Weak IV, large biases if slightly dependent on U (Small and Rosenbaum, 2008)
- Solution: strong IV!

Unobserved confounding (Endogeneity)



U:
-patients health
-expected gain from treatment
-centre context; physician characteristics
-level of risk aversion

IV to break endogeneity



Criteria for instrument

1. Predict D
2. ‘As if random’
3. Only effect on Y via D
4. SUTVA
5. Monotonicity

Near-far matching

Baiocchi et al, 2010; Zubizareta et al. 2013; Neuman et al, 2014

- Address bias from weak IV
- Study design approach to IV
- Replicate RCT within observational study
- Matching phase to prepare the data for analysis
- Find matched pairs ‘similar’ on observed covariate values (NEAR)
- Also strengthen IV to provide consistent estimates
- Within matched pairs find dissimilar individuals on levels of the IV
- Pairs of patients NEAR on covariates, FAR on IV
- IV highly predictive treatment receipt (criteria 1), independent of U (criteria 2)
- Well-designed RCT treatment and control regimens very different

Motivating example: evaluation of prompt ICU admission: Spotlight study

(Harris et al. 2015)

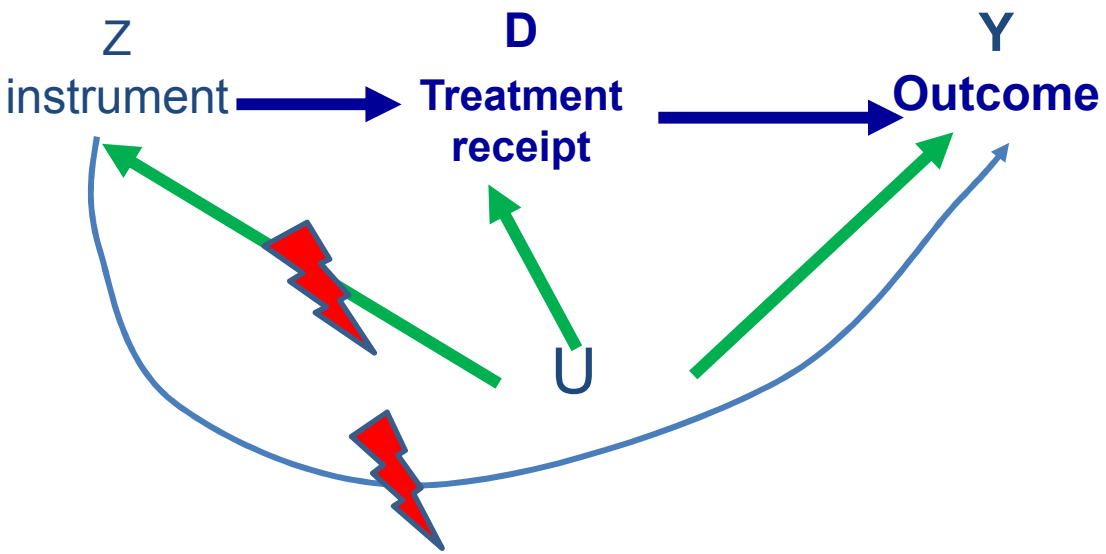
- Guidelines prompt (<4 hours) ICU admission for critically-ill patients
- ICU costly and beds limited, admission delayed or refused
- What is the effect of prompt admission on mortality?
- Precedent studies biased estimates inadequate design and analysis
- Focused entirely on ICU-admitted patients, selection on observables

- Spotlight study, 13,011 patients (48 hospitals) assessed for ICU admission
- Prompt ICU admission (<4 hours) vs not (later transfer or no transfer)
- 2,533 admitted promptly vs 10,478 not
- Naïve comparison and regression analysis, found prompt admission led to increased mortality

Spotlight: identification

- Prompt ICU admission (within 4 hours of assessment) versus not
- IV: bed availability in ICU at time of assessment,
- Median 4 beds available (range 0 to 19)
- Variation in beds available over time, and across centres.
- Argue exogenous especially within centres
- Exogenous factors (e.g. surgical lists) drive prompt admission versus not

Is Bed availability at ICU assessment (Z) an IV for prompt ICU admission (treatment, D)?



Criteria for instrument

1. Predict D: F statistic of 10
2. Independent of U
3. Only effect on Y via D

Covariate balance

comparison groups, median beds available

	'Few' beds available (N=6,114) Mean	'Many' beds available (N=6,897) Mean	Std. Diff	P-value
Age	64.95	65.40	-0.03	0.15
ICNARC score	15.15	15.01	0.02	0.26
Sepsis	0.61	0.61	-0.01	0.67
Peri-arrest	0.05	0.05	-0.03	0.12
Out of hours	0.38	0.33	0.09	0.00
Winter	0.32	0.20	0.28	0.00

Few: less than the median (4) beds available
Many: more than the median (4) beds available

Near-Far matching

Balanced observables, stronger IV

- Find matched pairs, near on covariates, far on IV
- Z_{ij} value of continuous instrument for patient j in possible pair i
- X_{ij} vector of observed pre-treatment covariates
- For ideal matched pair i_k and i_l

  $X_{ik}=X_{il}$

  Z_{ik}- Z_{il} would be large

- Within pairs, patients identical on observed covariates, one unit strongly encouraged to take treatment (prompt ICU care), other not

Near-far matching

- Penalty on matches that are 'near' on IV
- Distance penalty *p*

$$p = \begin{cases} (Z_{i1} - Z_{i2})^2 \, x \, c & \text{if } Z_{i1} - Z_{i2} < \Lambda \\ 0 & \text{otherwise} \end{cases}$$

Λ threshold pre-defined by analyst (a reverse caliper)

- Small matched distances for IV penalised, less likely to be matched
- Recognises tradeoff between strengthening IV and worsening balance
- Algorithm discards units hardest to match

Refined covariate balance
Pimental et al, 2016

- Spotlight study has 48 hospitals
- May be unobservable prognostic differences across hospitals, imbalanced between comparison groups
- Refined covariate balance can address this
- Designed to balance nested nominal covariates (e.g. hospitals)
- Balance no. patients per hospital for ‘few’ vs ‘more bed’ groups
- Requires more observations to be dropped, and changes estimand

Implementation

- Calculate Mahalanobis distance metric for all patients
- Match 1: strengthened the IV, but without refined balance
- Match 2: strengthened the IV, but with refined balance
- Report balance between comparison groups for pre-treatment covariates
- Report effect of intention to have ‘prompt’ ICU admission versus not
- Estimand is local treatment effect, subpopulation where the IV (no. of free beds) strongly encourages receipt of treatment (prompt ICU care)
- Report outcomes as 7 and 28 day mortality
- According to generalised effect ratios (Baiocchi et al, 2010)

Balance without vs with refined balance

	Strong IV 4,596 matched pairs			Strong IV, refined covariate balance, 2,048 matched pairs		
	'Few' beds available Mean	'Many' beds available Mean	Std. Diff	'Few' beds available Mean	'Many' beds available Mean	Std. Diff
ICU beds	1.68	7.64	2.91	1.56	7.05	3.07
Age	65.00	65.23	0.01	64.80	65.94	0.06
ICNARC score	15.23	15.07	0.02	15.59	15.57	0.00
Winter	0.21	0.21	0.00	0.27	0.27	0.00
Care level 0	0.08	0.04	0.17	0.11	0.11	0.02
variation hosp	30.25			5.55		
Total variation	32.44			6.23		

Outcomes: Estimated effect of prompt ICU admission on 7 and 28 day mortality

	Strong IV 4,596 matched pairs		Strong IV, refined balance 2,048 matched pairs	
Mortality	Mean [95% CI]	p-value	Mean [95% CI]	p-value
7 day	-0.031[-0.210, 0.144]	0.73	-0.252 [-0.642, 0.078]	0.132
28 day	-0.189[-0.638, 0.216]	0.132	-0.189 [-0.638, 0.216]	0.351

Interpretation and implications

- Design based approach, combine matching with IV
- Strengthened IV and maximise balance to reduce bias
- Extend matching algorithm accommodate design.
- Price is, local treatment effect for restricted sample, as in RCT
- Prompt ICU transfer reduction in mortality, not statistically significant
- Larger gains for more severe subgroups
- Design approach useful for evaluations that use natural experiments
- Could extend to other endpoints (e.g. cost)

Discussion and future work

- Much progress has been made in analysis of observational studies
- Less attention given to design
- Large data here to stay, increased opportunities for natural experiments in evaluation, especially with longitudinal linked data
- BUT increased scrutiny
- Essential to combine with appropriate study design and analysis
- May imply only identifying causal effects for subpopulations
- Consider reweighting treatment effects for alternative populations of prime policy-relevance (Hartman et al, 2013)
- Emphasis here on confounding, other problems warrant concurrent development.

References

- Baiocchi M et al (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. Journal American Statistical Association, 91,444-455
- Harris S (2015). Delay to admission to critical care and mortality among deteriorating ward patients in UK hospitals: a multicentre, prospective, observational study. Lancet 385, S40
- Keele L et al (2016). Stronger instruments and refined covariate balance in an observational study evaluating prompt admission to ICU. Working paper
- Kreif N et al. (2014). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. Health Services and Outcomes Research Methodology 13 (2-4), 174-202, 2013.
- O'Neill S et al. (2016). Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services Research and Outcomes Methodology* 16, 1-21
- Pimental SD et al (2015). Large, sparse, optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. Journal of the American Statistical Association, 110, 515-27.
- Sekhon JS (2011). Matching: multivariate and propensity score matching with automated balance search. *Journal of Statistical Software*. 42(7)
- Sekhon JS & Grieve R (2011). A Matching Method for Improving Covariate Balance in Cost-Effectiveness Analyses. *Health Econ* 21:695-714