Discussing and Displaying the Effectiveness of Causal Models for Observational Studies

Joint Statistical Meetings Luncheon Roundtable: August 4, 2003 from 12:30 to 1:50 PM
Renaissance Parc 55 Hotel: Barcelona Room I and II, Session 139: **ML-11**

Leader: Thomas E. Love, Ph. D., Case Western Reserve University
E-mail: thomaslove@cwru.edu Web: www.chrp.org/love

Abstract:
There has been rising interest in approaches designed to draw causal inferences from observational or quasi-experimental data in studies of health outcomes and policy. Causal interpretations of observed associations in such settings can be tricky, largely due to the problems of selection bias. Methodologies such as propensity scores and instrumental variables are increasingly part of the statistician's toolbox. Statisticians involved with health policy research need to communicate with an audience of physicians and other consumers, who are often skeptical of these techniques.

At this roundtable, we will discuss the role of statisticians in [1] arguing the need for appropriate methodologies for dealing with self-selection in observational studies; [2] producing effective displays for validating assumptions and documenting findings, and [3] describing an observational study's results, assumptions, caveats and conclusions accurately and usefully for a clinical or policy-oriented audience.

Arguing the Need for Dealing with Self-Selection in Observational Studies

Randomized experiments are the "gold standard" – they ensure that subjects receiving different exposures are comparable. Yet, we cannot always do experiments – exposures may be harmful, controlled by a systemic process that will not yield control, beyond reach legally or financially. We're frequently interested in phenomena that do not lend themselves to randomized trials. Such trials often have limited external validity as well – due in many cases to exclusion criteria, and other phenomena that limit our ability to study "entrenched practices".

In an observational study concerning exposures and their effects, the researcher does not control the assignment of exposures. Despite this, we want to be able to compare groups who "looked similar" prior to exposure assignment – thus, analytical adjustments are need to account for baseline differences in covariates. A study is biased if the exposed and unexposed groups differ in ways that matter for the outcomes of interest. We need to think hard about how exposure was determined.

Patients who receive an exposure are usually different from patients who don't receive it in important ways. We capture reasons behind exposure assignment in covariates, then adjust for covariate differences in estimating effects on outcomes.

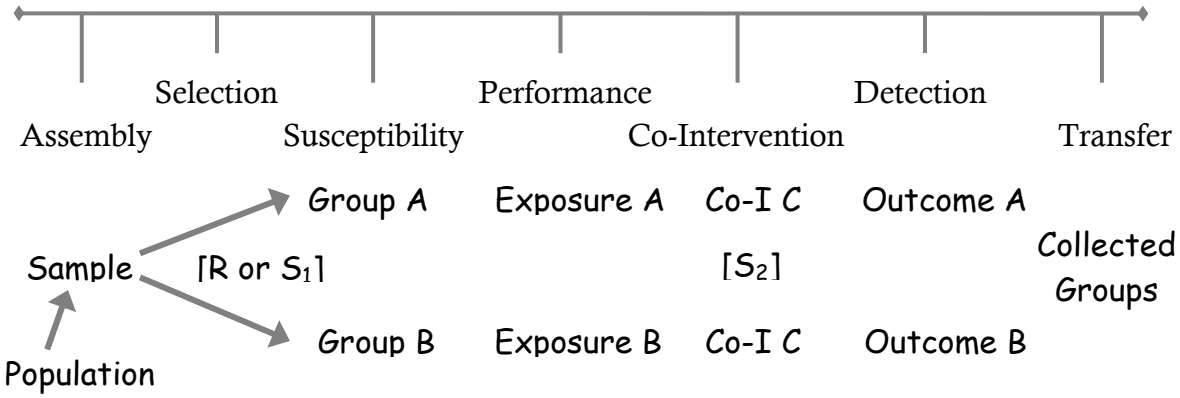
> "… the elucidation of causal relationships from observational studies must be shaped by knowledge (or assumptions) about how the data were generated; ***such assumptions are crucial to causal inference***."
>
> - Judea Pearl (2001) *HS&ORM*

Discussing and Displaying the Effectiveness of Causal Models for Observational Studies

Seven Key Aspects of Research Architecture: Judgments About Causation

(**Feinstein Model** for the evaluation of the scientific quality of cause-effect research, as modified by Neal V. Dawson, MD nvd@cwru.edu)

Goal: Comparability of the Groups Who Did and Did Not Receive the Exposure of Interest (Except for the Actual Receipt of Exposure)



1. **Distorted Assembly**: The sample should reflect the population to which results will be generalized. Application of specific inclusion/exclusion criteria will determine the pool of baseline characteristics of the sample.
2. **Selection Bias**: Bias can occur when subjects are selected to receive an exposure or co-intervention – especially when exposure is based on baseline covariates, and covariates are related to different likelihoods of outcome. Unmeasured covariates may or may not also be associated with measured characteristics.
3. **Susceptibility Bias / Case Mix / Severity**: Comparability of baseline characteristics of the exposure groups – are there importantly different expectations at baseline of the outcome of interest.
4. **Performance Bias**: How "well" do patients receive exposures (i.e. differences in dosage schedules, compliance rates, etc.)
5. **Co-Interventions** (Another Opportunity for Selection): Additional (medical) interventions beyond the exposure of interest that may influence the likelihood of achieving the outcome.
6. **Outcome Bias**: Process for determining the status of the outcome of interest in each group is applied unequally – differences in surveillance, diagnostic interpretation or testing, etc.
7. **Transfer Bias**: Members of the original or complete cohorts of A and B may be lost to dropout, intra-study exclusions, crossover, during statistical manipulations, etc.

"*Care in design and implementation will be rewarded with useful and clear study conclusions… Elaborate analytical methods will not salvage poor design or implementation of a study.*"

- National Academy of Sciences Report (Meyer and Feinberg 1992, p. 106)

Discussing and Displaying the Effectiveness of Causal Models for Observational Studies

The **standard ANCOVA procedure for mitigating selection bias**: (***Risk Adjustment***)

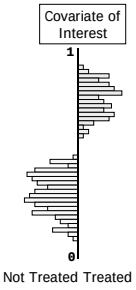
1. Capture as many important variables as possible and include them in the model, along with an indicator I_T or whether or not the subject received the treatment.
2. Identify the coefficient of the I_T indicator as a measure of the risk-adjusted "association" between receipt of treatment and the outcome.
3. Discuss and evaluate alternative explanations for the observed association other than "causality".

There is a clear need to …

> *"move beyond these informal techniques (that often perform well in the hands of 'master users') to established conceptual frameworks and "user-friendly" protocols."*
>
> –Ash, Normand, Duan (2001) *HS&ORM*

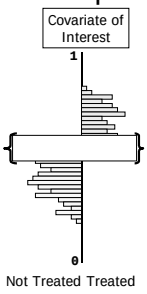
Validating Assumptions: The Issue of Overlap

How Much Overlap In The Covariates Do We Want?



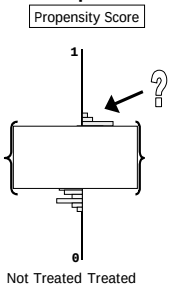
- If those who receive treatment don't overlap (in terms of covariates) with those who receive the control, we've got nothing to compare.
- Modeling, no matter how sophisticated, can't help us to develop information out of thin air.

What if Treated and Untreated Groups Overlap, but minimally?



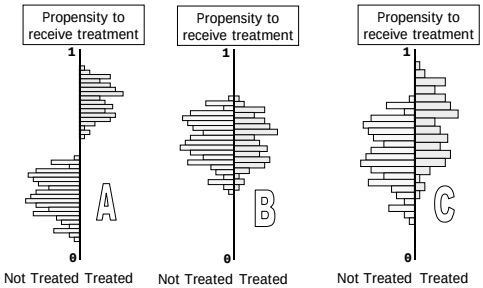
- No help.
- The information available to infer treatment effect will reside almost entirely in the few patients who overlap.
- Need to think hard about whether useful inferences will be possible.

What if Treated and Untreated Groups Don't Overlap Completely?



- Inferences for the causal effects of treatment on the subjects with no overlap cannot be drawn without heroic modeling assumptions.
- Usually, we'd exclude these treated subjects, and explain separately.

How Much Overlap In The Propensity Scores Do We Want?



Aspirin Use and Mortality Example from Gum et al. (2001) *JAMA*

Aspirin Use and Mortality

- 6174 consecutive adults undergoing stress echocardiography for evaluation of known or suspected coronary disease.
- 2310 (37%) were taking aspirin (treatment).
- Main Outcome: all-cause mortality
- Median follow-up: 3.1 years
- Univariate Analysis: 4.5% of aspirin patients died, and 4.5% of non-aspirin patients died…
- Unadjusted Hazard Ratio: 1.08 (0.85, 1.39)

Gum PA et al. (2001) *JAMA*, 1187-1194.

What Would Be Relevant Covariates for Adjustment?

- Demographics (Age, Sex)
- Cardiovascular risk factors
- Coronary disease history
- Use of other medications
- Ejection fraction
- Exercise capacity
- Heart rate recovery
- Echocardiographic ischemia

Result of adjusting for these factors: Aspirin use now associated with reduced mortality: Hazard Ratio: 0.67 95% CI: (.51, .87) p = .002

Baseline Characteristics According to Aspirin Use (before matching)

Variable	Aspirin* (n = 2310)	No Aspirin* (n = 3864)	P value
Age, years	62 (11)	56 (12)	< .001
Body mass index, kg/m^2	29 (5)	30 (7)	< .001
Ejection fraction, %	50 (9)	53 (7)	< .001
Resting heart rate, beats/min	74 (13)	79 (14)	< .001
Resting systolic BP, mm Hg	141 (21)	138 (20)	< .001
Resting diastolic BP, mm Hg	85 (11)	86 (11)	.04
Heart rate recovery, beats/min	28 (11)	30 (12)	< .001
Peak exercise cap., men (METs)	8.6 (2.4)	9.1 (2.6)	< .001
Peak exercise capacity, women	6.6 (2.0)	7.3 (2.1)	< .001

*Cells contain mean (SD)

Baseline Characteristics By Aspirin Use (in %) (before matching)

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	P value
Men	77.0	56.1	< .001
Clinical history: diabetes	16.8	11.2	< .001
hypertension	53.0	40.6	< .001
prior coronary artery disease	69.7	20.1	< .001
congestive heart failure	5.5	4.6	.12
Medication use: Beta-blocker	35.1	14.2	< .001
ACE inhibitor	13.0	11.4	< .001

- Baseline characteristics appear very dissimilar: 25 of 31 covariates have p < .001, 28 of 31 have p < .05.
- Aspirin user covariates indicate higher mortality risk.

Baseline Characteristics According to Aspirin Use (after matching)

Variable	Aspirin* (n = 1351)	No Aspirin* (n = 1351)	P value
Age, years	60 (11)	61 (11)	.16
Body mass index, kg/m^2	29 (6)	29 (6)	.83
Ejection fraction, %	51 (8)	51 (9)	.65
Resting heart rate, beats/min	77 (13)	76 (14)	.13
Resting systolic BP, mm Hg	141 (21)	141 (21)	.68
Resting diastolic BP, mm Hg	85 (11)	86 (11)	.57
Heart rate recovery, beats/min	28 (12)	28 (11)	.82
Peak exercise cap., men (METs)	8.7 (2.5)	8.3 (2.5)	**.01**
Peak exercise capacity, women	6.5 (2.0)	6.7 (2.0)	.13

*Cells contain mean (SD)

Baseline Characteristics By Aspirin Use [%] (after matching)

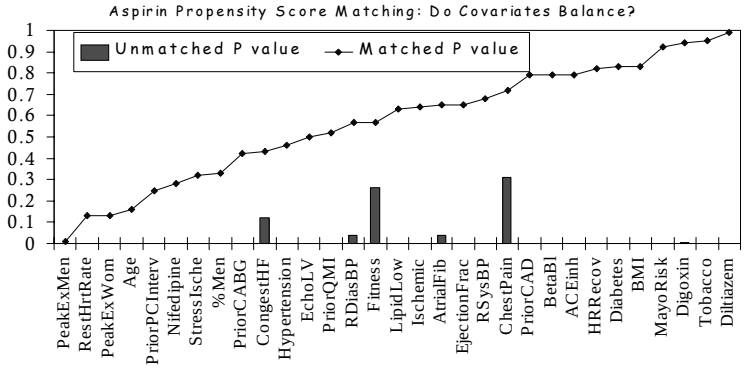
Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P value
Men	70.4	72.1	.33
Clinical history: diabetes	15.0	15.3	.83
hypertension	50.3	51.7	.46
prior coronary artery disease	48.3	48.8	.79
congestive heart failure	5.8	6.6	.43
Medication use: Beta-blocker	26.1	26.5	.79
ACE inhibitor	15.5	15.8	.79

- Baseline characteristics similar in matched users and non-users.
- 30 of 31 covariates show NS difference between matched users and non-users. [Peak exercise capacity for men is p = .01]

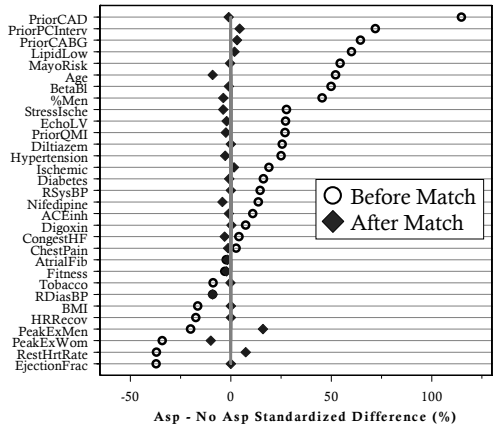
Displays for Validating Assumptions: Covariate Balance After Adjustment

Are The Covariates Balanced?
"P values Plot"

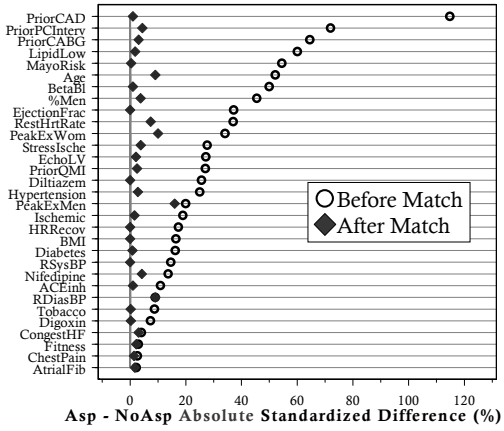
31 covariates provided in paper's Tables.



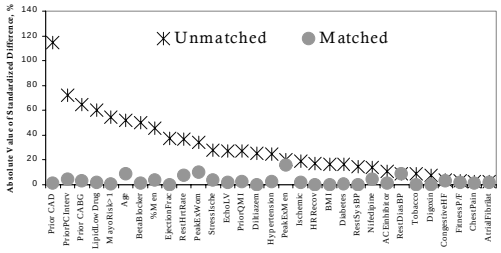
Covariate Balance for Aspirin Study



Absolute Standardized Differences



Covariate Balance for Aspirin Study

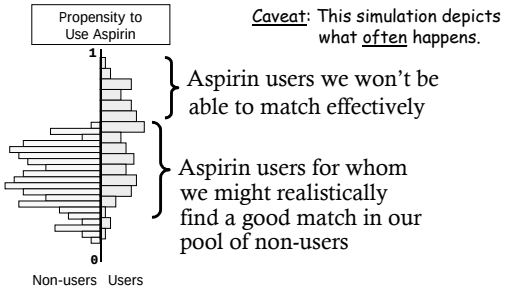


Absolute Value of Standardized Difference
plotted for 31 covariates (in Excel).

Discussing and Displaying the Effectiveness of Causal Models for Observational Studies

Who's Getting Matched Here?

Where Do The Propensity Scores Overlap?



Which Aspirin Users Get Matched?

- 652 of the 1351 matched aspirin users had had prior coronary artery disease (48.3%).
- 957 of the 959 unmatched aspirin users had had prior coronary artery disease (99.8%).

Variable	% of Matched	% of Unmatched	Stdzd Diff
Prior CAD	48.3	99.8	-145
Prior PCI	12.3	52.2	-95
Lipid-low th	20.8	51.5	-68
Prior CABG	18.6	45.7	-61
β-blocker	26.1	47.9	-46
Tobacco	11.9	7.3	+16

Describing the Results of an Observational Study Accurately and Usefully

Rosenbaum's Four Specific Suggestions (2002 book, p. 368)

- Design Observational Studies
  - Exert as much experimental control as possible, carefully consider the selection process, and anticipate hidden biases.
- Focus on Simple Comparisons
  - Increase impact of results on consumers
- Compare Subjects Who Looked Comparable Prior to Treatment
- Use Sensitivity Analyses to Inform Discussions of Hidden Bias Due to Unobserved Covariates
  - **Sensitivity analysis** asks how much hidden bias would need to be present to explain the differing outcomes in the exposed and control groups.

What Should Always Be Done in an Observational Study and Often Isn't

- Collect data so as to be able to model selection
- Demonstrate selection bias
- Ensure covariate overlap for comparability
- Evaluate covariate balance after adjustment
- Specify relevant post-adjustment population with care
- Model or estimate treatment effect in light of selection bias adjustments
- Estimate sensitivity of results to potential hidden biases

On Instrumental Variables vs. Propensity Methods vs. "Standard Risk Adjustment"

IV analysis has a long history in economics, where we are often facing "weak" data. The methods are attractive because they mirror RCT – instrument should adjust for both overt and hidden biases, and the resulting local average treatment effect estimates are sometimes more interesting than estimates from propensity models. It remains awfully difficult to justify the IV assumptions, although some specific examples (noncompliance in randomized trials) look well-suited to instruments. See

the Special Issue of HS&ORM described on page 10 of this handout for more details and comparisons (especially the Landrum/Ayanian and Posner et al. articles).

A Partial Bibliography

Articles I Have Used in A Short Course on Propensity Methods

1. Connors AF, Speroff T, Dawson NV, Thomas C, et al. (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients (with Editorial). *Journal of the American Medical Association*, 276: 889-897, 915-918. [Propensity scores in action: matching and multivariate adjustment.]
2. D'Agostino RB Jr. (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. [Readable introduction to matching, stratification and regression with propensity scores.]
3. Earle CC, Tsai JS, et al. (2001) Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology*, 19: 1064-1070. [Interesting combination of instrumental variables and propensity analyses.]
4. Gum PA, Thamilarasan M, Watanabe J, et al. (2001) Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease. *Journal of the American Medical Association*, 286(10)[Sep 12]: 1187-1194. Editorial: Radford MJ and Foody JM, pp. 1228-1230. [PS matching with nonproportional hazards models for survival analysis, the results show interesting impact of bias adjustment – editorial motivates observational studies vs. RCTs discussion, discusses cluster analysis and hierarchical modeling.]
5. Joffe MM and Rosenbaum PR (1999) Propensity scores. *American Journal of Epidemiology*, 150: 327-333. [Motivation for propensity score approaches, as well as a set of extensions – to case/control studies and to dose-response issues.]
6. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ et al. (2001) Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, 54: 387-398. [PS matching with calipers. Interesting discussion of covariate balance, leads to odds ratios, assessment of association through Mantel-Haenszel, survival rates, etc.]
7. Rosenbaum PR (1991) Discussing hidden bias in observational studies. *Annals of Internal Medicine*, 115: 901-905. [Introduction to sensitivity analysis, related issues.]
8. Rosenbaum PR and Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association*, 79: 516-524. [One of the seminal papers – presents subclassification on the propensity score.]
9. Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127: 757-763. [Introductory discussion, very readable.]

Relevant Books on Related Issues

10. **Rosenbaum PR (2002) *Observational Studies*, 2nd Edition. New York: Springer.**
11. Cochran WG (1983) *Planning and Analysis of Observational Studies* New York: Wiley.
12. Cook TD and Campbell DC (1979) *Quasi-Experimentation* Chicago: Rand McNally.
13. Elwood JM (1988) *Causal Relationships in Medicine* New York: Oxford Univ Press.
14. Pearl J (2000) *Causality: Models, Reasoning, Inference* New York: Cambridge Univ Press.
15. Shadish WR, Cook TD and Campbell DT (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* Boston: Houghlin-Mifflin.

Propensity Score Methodology – Development and Extensions

16. D'Agostino RB Jr and Rubin DB (2000) Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association [JASA] 95*, 749-759
17. Drake C and Fisher L (1995) Prognostic models and the propensity score. *International Journal of Epidemiology* 24: 183-187
18. Leon AC, Mueller TI et al. (2001) A dynamic adaptation of the propensity score adjustment for effectiveness analysis of ordinal doses of treatment. *Statistics in Medicine* 20: 1487-1498
19. Lu B, Zanutto E, Hornik R and Rosenbaum PR (2002) Matching with doses in an observational study of a media campaign against drug abuse. *JASA 96*, 1245-1253.
20. Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41-55
21. Rosenbaum PR (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment, *Journal of the Royal Statistical Society, Ser. A* 147, 656-666.
22. Rosenbaum PR and Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*, 33-38.
23. Rosenbaum PR (1987) Model-based direct adjustment. *JASA 82*, 387-394.
24. Rubin DB and Thomas N (1996) Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52: 249-264.
25. Rubin DB and Thomas N (2000) Combining propensity score matching with additional adjustments for prognostic covariates. *JASA 95*, 573-585.
26. Wang J, Donnan PT et al. (2001) The multiple propensity score for analysis of dose-response relationships in drug safety studies, *Pharmacoepidemiology and Drug Safety 10,* 105-111.

Instrumental Variables and Related Ideas

27. Angrist JD, Imbens GW and Rubin DB (1996) Identification of causal effects using instrumental variables (with Comments). *JASA 91*, 444-472. [Key IV paper for statisticians.]
28. Greenland S (2000) An introduction to instrumental variables for epidemiologists, *International Journal of Epidemiology 29*, 722-729.
29. Lavori PW, Dawson R, Mueller TB (1994) Causal estimation of time-varying treatment effects in observational studies - Application to depressive disorder, *Statistics in Medicine 13*, 1089-1100.
30. McClellan M, McNeil BJ, Newhouse JP (1994) Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables, *Journal of the American Medical Association 272*, 859-866.
31. Newhouse JP, McClellan M (1998) Econometrics in outcomes research: the use of instrumental variables, *Annual Review of Public Health 19*, 17-34.

Sensitivity Analysis

32. Greenland S (1996) Basic methods for sensitivity analysis of biases, *International Journal of Epidemiology*, *25*, 1107-1116.
33. Lin DY, Psaty BM and Kronmal RA (1998) Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics*, *54,* 948-963.
34. Rosenbaum PR (1999) Choice as an Alternative to Control in Observational Studies (with Comments), *Statistical Science*, *14*, 259-304.
35. Rosenbaum PR and Rubin DB (1983) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society, Series B*, *45,* 212-218.

Causation, Confounding and The Role of Observational Studies

36. Abel U, Koch A (1999) The role of randomization in clinical studies: Myths and beliefs, *Journal of Clinical Epidemiology 52,* 487-497.
37. Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics 24*, 295-313.
38. Cochran WG (1965) The planning of observational studies of human populations (with discussion), *Journal of the Royal Statistical Society, Series A 128*, 234-255.
39. Copas JB and Li HG (1997) Inference for Non-Random Samples (with discussion), *Journal of the Royal Statistical Society, Series B 59,* 55-95.
40. Greenland S and Morgenstern H (2001) Confounding in health research, *Annual Review of Public Health 22*, 189-212.
41. Greenland S (2000) Causal Analysis in the Health Sciences, *JASA 95*, 286-289.

42. Kraemer HC, Stice E et al. (2001) How do risk factors work together? Mediators, moderators and independent, overlapping, and proxy risk factors, *American Journal of Psychiatry 158,* 848-856.
43. Little RJA and Rubin DB (2000) Causal effects in clinical and epidemiological studies via potential outcomes. *Annual Review of Public Health 21*, 121-145.
44. Rosenbaum PR (1984) From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment, *JASA 79*, 41-48.
45. Rosenbaum PR (1987) The role of a second control group in an observational study (with Discussion.) *Statistical Science, 2,* 292-316.
46. Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics 47,* 1213-1234.

Special Issue of Health Services & Outcomes Research Methodology

An International Journal devoted to Quantitative Methods for the Study of Utilization, Quality, Cost and Outcomes of Health Care - **December 2001, Volume 2, Issue 3-4**
Available Online [Free PDF] at http://www.kluweronline.com/issn/1387-3741