Eron LD, Huesmann LR, Lefkowitz MM, Walder LO. Does television violence cause aggression? Am Psychol. 1972;27:253–63. PubMed

Money Supply

Granger Causality. Nobel 2003. Complete Granger

Relationships--and the Lack Thereof--Between Economic Time Series, with Special Reference to Money and Interest Rates. David A. Pierce *Journal of the American Statistical Association*, Vol. 72, No. 357. (Mar., 1977), pp. 11-26. Jstor

Additional Resources

Reciprocal effects: Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, 88, 245-258. APA site version

Structural Equation Modeling With the sem Package in R John Fox STRUCTURAL EQUATION MODELING,13(3),465- 486 Jox Fox home page

Week 6 Review Questions

Question 1. Grouping and multilevel regressions

Illustrate relations among individual level (ignoring groups) group-level, and relative standing regression results.

Part I groups formed on X

Create 200 individual level observations on X and Y having correlation around .65.

I started with x values 1:200 (simple integers) for convenience, but you can be fancier.

Do an individual level Y on X regression (i.e. "total, ignoring groups which don't exist yet).

Group these 200 individuals into 10 groups of size 20 on the basis of the X-values (i.e. group 1 contains the individuals with the smallest 20 X-values, group 10 contains the individuals with the largest 20 X-values). So within-groups will be as homogeneous as possible on X, and between group differences on X will be largest.

Do a regression on group means (between groups regression) these may be classroom means for example, and you may not have individual level data.

Get a relative standing measure: individual score minus group mean for each individual.

Do a relative standing regression

Now do the multiple regression analyses (class handouts; Burstein, Deleuuw & Kreft)

1. "context" Y on X and X-bar (X-bar is an attribute of each individual)
2. "Cronbach" (Kreft's term) Y on X minus X-bar and X-bar (predictors uncorrelated)

Demonstrate the coefficients match the basic relations shown in lecture

Part II groups formed independent of X (random)

Repeat the analyses of Part I using a different (as different as can be) mechanism for assigning individuals to groups. Form the 10 groups of size 20 at random, making the groups heterogeneous on X within group and similar between groups.

Solution for question 1

Question 2. Contextual Effects Coefficient

Use the regression recursion relation from week 4 to show that the contextual effects coefficient defined in week 6 handouts is equal as stated in the handouts (and literature) to the between groups slope minus the within-pooled slope.

Solution for question 2

Question 3. Simplified version of HSB analysis

The ubiquitous analyses of the HSB data use a level 2 model, with meanses as a covariate in addition to the 'group treatment' indicator sector (P/C).

For intro instruction use of these multilevel methods for comparing 'effects' of Public vs Catholic, it would be cleaner just to do a 't-test' in the level 2 model-- i.e. the only predictor of level and gradient being sector.

Try out that simpler model and compare with standard analysis. Note that the side-by-side boxplots are still relevant for this reduced model, as the boxplots only relect the Level 1 specifications.

Solution for question 3

Question 4. Enrichment problem (better to spend time on HSB analyses etc)

Ecological fallacy: Is Radon good for you?

Treat this as an extended example of ecological bias.

At one time I went through the Robbins paper in class...

Solutions show you data generation procedures and illustrate the sometimes very large effects of aggregation bias. If the topic interests read through the G-R paper to see the point.

Consider the artificial data example described in Ex 3 p.750 Greenland and Robbins American Journal of Epidemiology Vol. 139, No. 8: 747-760 Ecologic Studies—Biases, Misconceptions, and Counterexamples (article linked on class page, week 6 under additional resources)

intro their Example 3

Suppose that our study data are limited to regional values of mean radon, mean smoking (in packs per day), and lung-cancer rates among males aged 70-74 years, for 41 regions indexed by r = 0, . . . , 40.

follow their example set up and create your own artificial data example and produce the regression function and plot in their figure 1 for the effect of radon levels on lung cancer rates

from G&R you are demonstrating the ecological fallacy because "the regressions yield an inverse association of radon and lung cancer, despite the fact that radon is a positive risk factor in the underlying model used to generate the data,"

"Even though the lung-cancer rates show the strong upward relation to smoking one would expect from model 1, and the ecologic correlation between radon and smoking is only 0.01, there is a significant negative ecologic association of radon with lung cancer rates."

Solution for question 4

Question 5. Simultaneous effects.

For the Duncan Haller Portes occupational aspiration example from class handout (cf Fox Soc Meth 1979 paper) replicate the 2SLS (IV) analysis of this non-recursive model from the class handout.

Extra item: Can you fit a model which adds a path from Friend's family SES to respondents occupational aspiration?

Solution for question 5

Week7

1. Matching Methods for Observational Data: Part I

Lecture topics
0. Review: Matching for increased precision, Randomized block designs (see Review Questions) package `blockTools`
1. Traditional matching methods: subclassification, pair matching. Case-control studies.
handout for smoking ex, Cochran subclassification
2. Modern Implementations of matching methods The advent/onslaught of propensity score matching methodology for treatment-control comparisons
optmatch exs, nuclear plants, gender ascii version for some Ben Hansen matching exs using MatchIt/optmatch
propensity score intro checking balance, aspirin ex

Primary Readings
Case-control studies: Case-control overview from Encyclopedia of Public Health
Non-technical matching overviews: Donald Rubin Nonrandomized Comparative Clinical Studies another version,[Lane library from campus] *Annals of Internal Medicine,* 1997, 15 October 1997, Vol. 127. No. 8_Part_2
Cochran's smoking, subclassification and Rubin's Breast Cancer example also discussed in Rubin "Design Trumps Analysis" Rubin paper . also set of slides
Another Rubin overview of matching, Matching Methods for Causal Inference Elizabeth Stuart Donald Rubin [does Lalonde example]
Joffe, Marshall M. and Paul R. Rosenbaum. 1999. "Invited Commentary: Propensity Scores." American Journal of Epidemiology 150(4):327-33.
Rosenbaum and Rubin, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, JASA 79[387], September 1984, 516-524. JStor [one of the original technical papers]

Matching Research Examples
Aspirin Pair Matching
Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. Gum PA1, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. JAMA. 2001 Sep 12;286(10):1187-94.
SAT Coaching, Full Matching
Optmatch application paper: Hansen, Ben B. Full matching in an observational study of coaching for the SAT.(Scholastic Assessment Test) *Journal of the American Statistical Association*; 9/1/2004;
Coronary Artery Disease
Rosenbaum and Rubin, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, JASA 79[387], September 1984, 516-524. JStor
Breastfeeding and Propensity Scores
Breastfeeding May Not Lead to Smarter Preschoolers Breastfeeding does NOT boost a baby's IQ: Nourishing infants the natural way only makes them less hyper
Breast-feeding study sheds light on benefits for babies
Publication: Breastfeeding, Cognitive and Noncognitive Development in Early Childhood: A Population Study. Lisa-Christine Girard, Orla Doyle, Richard E. Tremblay. PEDIATRICS Volume 1 39, number 4 , April 2017.

Additional resources
Talks and tutorials
Strategies for Using Propensity Scores Well. A Workshop given by Thomas E. Love, Ph. D., Case Western Reserve University Love workshop ASA
A broad review of matching and bias-reduction methods. Opiates for the Matches: Matching Methods for Causal Inference Jasjeet S. Sekhon. *Annual Review of Political Science* 2009
UNC, Chapel Hill Social Work: Introduction to Propensity Score Matching: A Review and Illustration Propensity Score Matching: A New Device for Program Evaluation UNC, Chapel Hill Social Work 2004 flash version
An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies Peter C. Austin Multivariate Behav Res. 2011 May; 46(3): 399-424.
Methods to assess intended effects of drug treatment in observational studies are reviewed *Journal of Clinical Epidemiology* 57(2004)1223-1231 [an overview of many of past weeks topics]
Average causal effects from nonrandomized studies: A practical guide and simulated example. Schafer, Joseph L.; Kang, Joseph Psychological Methods, Vol 13(4), Dec 2008, 279-313.
A Primer for Applying Propensity-Score Matching Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank
Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group Statist. Med. 17, 2265-2281 (1998)

R packages and examples:
1. Ben Hansen (local hero) optmatch manual R News Oct 2007 Hansen presentation: Flexible, Optimal Matching for Comparative Studies Using the optmatch package
Optmatch application paper: Full matching in an observational study of coaching for the SAT.(Scholastic Assessment Test) *Journal of the American Statistical Association*; 9/1/2004; Hansen, Ben B.
Additional exercises (checking balance) using the nuclearplants data (class handout ex) from Mark Fredrickson here
2. MatchIt: Nonparametric Preprocessing for Parametric Casual Inference Daniel Ho, Kosuke Imai, Gary King, Elizabeth Stuart MatchIt provides a wrapper that can call optmatch or Sekhon's genetic matching]
JSS May 2011 exposition: MatchIt: Nonparametric Preprocessing for Parametric Causal Inference more R-fun from Gary King, WhatIf: Software for Evaluating Counterfactuals
Another application (including matchit): Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference Jake Bowers and Ben Hansen. Data archive and computing resources for the New Haven get-out-the-vote
Also:
3. Multivariate and Propensity Score Matching Software for Causal Inference Jasjeet S. Sekhon

Propensity etc Original Technical Publications [jstor links]
Rosenbaum, P. R. And D. B. Rubin, 1983, The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika 70[1], April 1983, 41-55. JStor
P. Rosenbaum, Chapters 2 and 3 (on exact inference for treatment effects) in Observational Studies, New York: Springer, 1995.
Dropping out of High School in the United States: An Observational Study Paul R. Rosenbaum *Journal of Educational Statistics*, Vol. 11, No. 3. (Autumn, 1986), pp. 207-224. Jstor
Paul R. Rosenbaum; Donald B. Rubin. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score" *The American Statistician*, Vol. 39, No. 1. (Feb., 1985), pp. 33-38 JStor Danish downers example
D. Rubin, Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies, Statistical Science 5[4], November 1990, 472-480. JStor
Rubin, D. B., 1974, Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Journal of Educational Psychology, 66, 688-701.
Rubin, D. B., 1978, Bayesian Inference for Causal Effects: The Role of Randomization," Annals of Statistics 6[1], January 1978, 34-58. JStor

Case-control studies
Case-control overview (shown in class) from Encyclopedia of Public Health

Breslow NE. Statistics in epidemiology: the case-control study.J Am Stat Assoc. 1996 Mar;91(433):14-28
Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma JNCI: Journal of the National Cancer Institute, Volume 98, Issue 1, 4 January 2006, Pages 72-75,
Smoking and Lung Cancer in Chap 18 of HSAUR3 (Handbook of Statistical Analysis Using R). Also driving and backpain data in Chap 7 HSAUR2
Some R-packages and resources: SensitivityCaseControl: Sensitivity Analysis for Case-Control Studies; multipleNCC: Inverse Probability Weighting of Nested Case-Control Data; Two-phase designs in epidemiology (Thomas Lumley) ; Exact McNemar's Test and Matching Confidence Intervals

Weeks 7 and 8 Review Questions

Randomized Blocks, Experimental Designs

Question 1. Matching and Paired t-test example from lecture
(Stat 141 exam problem (circa 2005))
An experiment on treating depression by Imipramine, an anti- depressant drug, employed a matched-pairs design. A total of 60 patients were paired on a combination of age, sex, and time of entry in study to form 30 matched pairs. That is, each pair consisted of patients who entered the study within a month of each other, were of the same sex and were similar in age. One member of each pair was randomly assigned to receive Imipramine and the other to receive a placebo. The outcome measure was the score on the Hamilton rating scale for depression (higher score = more severe depression) after 5 weeks of treatment.
The file http://web.stanford.edu/~rag/stat209/depressdata contains the outcome scores for each of the 30 pairs (Imipramine vs Placebo).
a. Carry out a statistical test of the equality of treatment outcomes. That is, test null hypothesis that Imipramine and Placebo produce equivalent outcomes versus a non-directional alternative. Use Type 1 error rate .05. State the result of the statistical test.
b. Pretend that an erstwhile graduate assistant lost all records of the matched pairs before the data analysis could be completed. Consequently, all the investigator has available is the 30 scores for the patients receiving Imipramine and the 30 scores for the patients receiving Placebo (but not the information on the matching). Carry out a statistical test of the hypothesis in part a using the available information. Is the result of the test the same? Explain why or why not.
c. Regard part (b) as a bad dream and return to the data set with full matching information. But now you are told that the differences between Hamilton scale scores shouldn't be regarded as having numerical value. Comparing two Hamilton scores only indicates relative standing, that is which of the two patients in the matched pair is showing greater symptoms of depression. Under that limitation of the data carry out an appropriate statistical test of the hypothesis in part (a). Explain why the result is the same or different from the result in part (a).

Solution for question 1

Question 2. Matching to increase precision: Factorial Randomized blocks designs
Example from lecture, Neter-Wasserman problem DENTAL PAIN.
An anesthesiologist made a comparative study of the effects of acupuncture and codiene on postoperative dental pain in male subjects. The four treatments were (1) placebo treatment-- a sugar capsule and two inactive acupuncture points, (2) codiene treatment only--a codeine capsule and two inactive acupuncture points; (3) acupucture only--a sugar capsule and two active acupuncture points (4) both codeine and acupuncture. These 4 conditions have a 2x2 factorial structure.
Thirty-two subjects were grouped into 8 blocks of four according to an initial evaluation of their level of pain tolerance. The subjects in each block were then randomly assigned to the 4 treatments. Pain relief scores were obtained 2 hours after dental treatment. Data were collected on a double-blind basis.
Data in file: http://statweb.stanford.edu/~rag/stat209/dental.dat
c1 is pain relief score (higher means more pain relief); c2 is block; c3 is codiene; c4 is acupuncture--for c3 and c4, 1=no.
a. obtain cell means for the 2x2 factorial design
b. carry out the randomized blocks analysis of variance, factors are Block, main effects for Codeine Acup and interaction term Codeine*Acup
c. Give a measure for the relative efficiency of the blocking on pain tolerance--how much better in terms of precision or number of subjects needed is the analysis using blockings versus a 2x2 factorial design design that ignores pain tolerance?

Solution for question 2

Matching and propensity score methods, Observational Studies

Question 3.
Recreate the matching demonstration for Ben Hansen's "gender equity" example (done in the week 7 class handout, posted not hard copy), an example of optimal full matching. Only one matching variable. this is Example 2 in Hansen's talk, about p.48 in the linked pdf here's the data in cut-and-paste form

```
> geneq
  Grant gender
1   5.7      W
2   4.0      W
3   3.4      W
4   3.1      W
5   5.5      M
6   5.3      M
7   4.9      M
8   4.9      M
9   3.9      M
```

Solution for question 3

Question 4. Multivariate matching
The example shown in lecture, from anderson et al
Example 6.5 Multivariate caliper matching: Consider a hypothetical study comparing two therapies effective in reducing blood pressure, where the investigators want to match on three variables: previously measured diastolic blood pressure (DPB), age, and sex. Such confounding variables can be divided into two types: categorical variables, such as sex, for which the investigators may insist on a perfect match (e = 0); and numerical variables, such as age and blood pressure, which require a specific value of the caliper tolerances. Let the blood pressure tolerance be specified as 5 mm Hg and the age tolerance as 5 years. The data contains measurements of these three confounding variables. (The subjects are grouped by sex to make it easier to follow the example.)
Data with columns DBP age sex and Grp (Treatment Group or Comparison Reservoir) http://statweb.stanford.edu/~rag/stat209/matchex.dat

```
Table 6.6 Hypothetical Measurements on Confounding Variables
Treatment Group                          Comparison Reservoir
Subject Diastolic Blood                Subject  Diastolic Blood
Number  Pressure (mm Hg) Age Sex       Number    Pressure (mm Hg)   Age Sex
1          94            39  F          1             80          35  F
2          108           56  F          2             120         37  F
3          100           50  F          3             85          50  F
4          92            42  F          4             90          41  F
5          65            45  M          5             90          47  F
6          90            37  M          6             90          56  F
```

matching / blocking to improve precision

27.9. **Dental pain.** An anesthesiologist made a comparative study of the effects of acupuncture and codeine on postoperative dental pain in male subjects. The four treatments were: (1) placebo treatment—a sugar capsule and two inactive acupuncture points (A_1B_1), (2) codeine treatment only—a codeine capsule and two inactive acupuncture points (A_2B_1), (3) acupuncture treatment only—a sugar capsule and two active acupuncture points (A_1B_2), and (4) codeine and acupuncture treatment—a codeine capsule and two

HW8, part 1

active acupuncture points (A_2B_2). Thirty-two subjects were grouped into eight blocks of four according to an initial evaluation of their level of pain tolerance. The subjects in each block were then randomly assigned to the four treatments. Pain relief scores were obtained for all subjects two hours after dental treatment. Data were collected on a double-blind basis. The data on pain relief scores follow (the higher the pain relief score, the more effective the treatment).

Block i		Treatment (j, k)			
		A_1B_1	A_2B_1	A_1B_2	A_2B_2
1	(Lowest)	0.0	.5	.6	1.2
2		.3	.6	.7	1.3
...		...	...	...	...
7		1.0	1.8	1.7	2.1
8	(Highest)	1.2	1.7	1.6	2.4

Fleiss 5.1; multivariate matching, paired t-test

Table 5.1. Results of a study comparing Imipramine with placebo for the treatment of depression, with patients paired (blocked) by time of enrollment, sex, and age

Pair	Imipramine	Placebo	Difference	Pair	Imipramine	Placebo	Difference
1	6	4	2	16	6	8	−2
2	4	7	−3	17	10	10	0
3	6	12	−6	18	3	9	−6
4	7	10	−3	19	5	8	−3
5	5	2	3	20	4	5	−1
6	6	11	−5	21	6	8	−2
7	8	9	−1	22	7	7	0
8	7	5	2	23	5	6	−1
9	8	11	−3	24	6	9	−3
10	3	8	−5	25	3	3	0
11	9	7	2	26	10	5	5
12	4	6	−2	27	5	11	−6
13	8	8	0	28	4	7	−3
14	11	9	2	29	4	3	1
15	12	9	3	30	7	10	−3
				Mean	6.3000	7.5667	−1.2667
				sd	2.3947	2.5955	2.9235

data in depress.dat

the same sex, and were similar in age. For the pairs of females, similarity meant ages that were no more than 10 years apart. Because depression is rarer among males, similarity in age for pairs of males meant ages that were no more than 20 years apart; any tighter criterion would have resulted in many males not being pairable with another.

One member of each pair was randomly assigned to receive Imipramine and the other to receive placebo. The values in the table are scores on the Hamilton rating scale for depression (Hamilton, 1960) after five weeks of treatment (the higher the score, the more severe the depression). To appreciate the different statistical properties of a parallel groups and a randomized blocks design, consider the standard error of the estimated difference in means between the two treatments. The estimated standard error of the mean difference, $\bar{d}$, is

$$\text{se}(\bar{d}) = \frac{2.9235}{\sqrt{30}} = 0.53. \qquad (5.1)$$

Had the 60 patients not been paired, but studied in a parallel groups experiment with 30 patients per group, the standard error would be

.64

Package 'blockTools'

February 19, 2015

Type Package

Title Block, Assign, and Diagnose Potential Interference in Randomized Experiments

Version 0.6-2

Date 2015-01-08

Author Ryan T. Moore and Keith Schnakenberg

Maintainer Ryan T. Moore <rtm@american.edu>

Imports MASS

Suggests nbpMatching, RItools, xtable

Description Blocks units into experimental blocks, with one unit per treatment condition, by creating a measure of multivariate distance between all possible pairs of units. Maximum, minimum, or an allowable range of differences between units on one variable can be set. Randomly assign units to treatment conditions. Diagnose potential interference between units assigned to different treatment conditions. Write outputs to .tex and .csv files.

License GPL (>= 2) | file LICENSE

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-01-09 06:24:26

R topics documented:

`blockTools-package` *Block, Randomly Assign, and Diagnose Potential Interference in Randomized Experiments*

Description

Block units into experimental blocks, with one unit per treatment condition, by creating a measure of multivariate distance between all possible pairs of units. Maximum, minimum, or an allowable range of differences between units on one variable can be set. Randomly assign units to treatment conditions. Diagnose potential interference problems between units assigned to different treatment conditions. Write outputs to .tex and .csv files.

Details

Package:	blockTools
Type:	Package
Version:	0.6-2
Date:	2015-01-08
License:	GPL (>=2)

Given raw data, `block` creates experimental blocks, `assignment` assigns units to treatment conditions, `diagnose` detects possible interference problems, and `outTeX` and `outCSV` write block or assignment output objects to a set of .tex and .csv files, respectively. In sequential experiments, `seqblock` assigns units to treatment conditions.

Author(s)

Ryan T. Moore <rtm@american.edu> and Keith Schnakenberg <keith.schnakenberg@gmail.com>

Maintainer: Ryan T. Moore <rtm@american.edu>

References

http://ryantmoore.com/software.blockTools.htm

Examples

```
data(x100)

## block
out <- block(x100, groups = "g", n.tr = 2, id.vars = c("id"), block.vars
             = c("b1", "b2"), algorithm="optGreedy", distance =
             "mahalanobis", level.two = FALSE, valid.var = "b1",
             valid.range = c(0,500), verbose = TRUE)
## assign
assg <- assignment(out, seed = 123)
```

Author(s)

Ryan T. Moore

References

Hansen, Ben B. and Jake Bowers. 2008. "Covariate balance in simple, stratified and clustered comparative studies". *Statistical Science* 23(2):219–236.

Bowers, Jake and Mark Fredrickson and Ben Hansen. 2010. "RItools:Randomization Inference Tools". R package version 0.1-11.

Moore, Ryan T. 2012. "Multivariate Continuous Blocking to Improve Political Science Experiments". *Political Analysis*, 20(4):460–479, Autumn.

See Also

assignment

Examples

```
data(x100)
b <- block(x100, groups = "g", id.vars = "id", block.vars = c("b1", "b2"))
a <- assignment(b)
axb <- assg2xBalance(a, x100, id.var = "id", bal.vars = c("b1", "b2"))
axb
## axb is a list with 4 elements (one for each of 3 groups, plus one for 'Overall')
```

`assignment` *Randomly assign blocked units to treatment conditions*

Description

Using an output object from `block`, assign elements of each row to treatment condition columns. Each element is equally likely to be assigned to each column.

Usage

```
assignment(block.obj, seed = NULL, namesCol = NULL)
```

Arguments

`block.obj`	an output object from `block`, or a user-specified block object.
`seed`	a user-specified random seed.
`namesCol`	an optional vector of column names for the output table.

Stat209/Ed260 D Rogosa 2/26/18

Assignment 8. Matching review: Randomized Block Designs

Problem 1. Matching and Paired t-test
Example from lecture
Stat 141 exam problem (circa 2005)

An experiment on treating depression by Imipramine, an anti-depressant drug, employed a matched-pairs design. A total of 60 patients were paired on a combination of age, sex, and time of entry in study to form 30 matched pairs. That is, each pair consisted of patients who entered the study within a month of each other, were of the same sex and were similar in age. One member of each pair was randomly assigned to receive Imipramine and the other to receive a placebo. The outcome measure was the score on the Hamilton rating scale for depression (higher score = more severe depression) after 5 weeks of treatment.

The file http://web.stanford.edu/~rag/stat209/depressdata contains the outcome scores for each of the 30 pairs (Imipramine vs Placebo).

a. Carry out a statistical test of the equality of treatment outcomes. That is, test null hypothesis that Imipramine and Placebo produce equivalent outcomes versus a non-directional alternative. Use Type 1 error rate .05. State the result of the statistical test.

b. Pretend that an erstwhile graduate assistant lost all records of the matched pairs before the data analysis could be completed. Consequently, all the investigator has available is the 30 scores for the patients receiving Imipramine and the 30 scores for the patients receiving Placebo (but not the information on the matching). Carry out a statistical test of the hypothesis in part a using the available information. Is the result of the test the same? Explain why or why not.

c. Regard part (b) as a bad dream and return to the data set with full matching information. But now you are told that the differences between Hamilton scale scores shouldn't be regarded as having numerical value. Comparing two Hamilton scores only indicates relative standing, that is which of the two patients in the matched pair is showing greater symptoms of depression. Under that limitation of the data carry out an appropriate statistical test of the hypothesis in part (a). Explain why the result is the same or different from the result in part (a).

Problem 2. background & review
Matching to increase precision: Factorial Randomized blocks designs
Example from lecture
From Neter-Wasserman problem DENTAL PAIN.

An anesthesiologist made a comparative study of the effects of acupuncture and codiene on postoperative dental pain in male subjects. The four treatments were (1) placebo treatment-- a sugar capsule and two inactive acupuncture points, (2) codiene treatment only--a codeine capsule and two inactive acupuncture points; (3) acupucture only--a sugar capsule and two active acupuncture points (4) both codeine and acupuncture. These 4 conditions have a 2x2 factorial structure.
Thirty-two subjects were grouped into 8 blocks

of four according to an initial evaluation of
their level of pain tolerance. The subjects in
each block were then randomly assigned to the 4
treatments. Pain relief scores were obtained 2
hours after dental treatment. Data were
collected on a double-blind basis.
Data in file
http://www-stat.stanford.edu/~rag/stat209/dental.dat
c1 is pain relief score (higher
means more pain relief), c2 is block c3 is
codiene c4 is acupuncture--for c3 and c4, 1=no.

a. obtain cell means for the 2x2 factorial design
b. carry out the randomized blocks analysis of variance,
factors are Block, main effects for Codeine Acup and
interaction term Codeine*Acup,
c. Give a measure for the relative efficiency of the blocking
on pain tolerance--how much better in terms of precision
or number of subjects needed is the analysis using blockings
versus a 2x2 factorial design design that ignores pain tolerance?

==========
end homework 8 part 1

LIFE & CULTURE | JANUARY 17, 2012

Bypass Beats Band for Weight Loss

By JENNIFER CORBETT DOOREN

A study of two popular surgical procedures to treat morbidly obese patients shows gastric bypass is associated with faster and more sustained weight loss than gastric banding.

Weight loss was faster, greater and remained "significantly better" six years after gastric bypass compared with patients who received a gastric band, according to researchers.

The study, which involved more than 400 patients in Switzerland, is one of the longest studies of the two common procedures in the U.S. that limit the amount of food the stomach can hold. It was published online Monday in the Archives of Surgery.

Weight loss was measured by looking at group changes in body mass index at various times after surgery. BMI is a measure that estimates body fat by using a person's height and weight. People with BMIs of 30 or higher are considered obese.

Study participants started out with an average BMI of about 43. After a year, the average BMI in the bypass group fell below 30, while those receiving a gastric band had a BMI of about 34. Researchers said maximum weight loss was achieved after an average of three years for the gastric-band patients compared with 18 months for the bypass group.

Total cholesterol remained unchanged in patients who underwent gastric banding but decreased in patients who underwent gastric bypass.

The study looked at what are considered treatment failures, which was measured by a reversal of the procedures, or patients who had a BMI of 35 six years after surgery. The failure rate for gastric banding was 48.3% compared with 12.3% for bypass.

Gastric banding involves the placements of a band around the top part of the stomach to create a small pouch. In gastric-bypass surgery, surgeons reduce the size of the stomach and the smaller stomach is then attached to the middle of the small intestine, bypassing a section of the intestine and thereby limiting the absorption of calories. About 200,000 Americans undergo surgical procedures to shrink their stomachs each year.

The study involved 442 patients who were operated on between March 1998 and May 2005. Half of the patients received a gastric band while the other half underwent the Roux-en-Y procedure, a common gastric-bypass procedure in the U.S. Patients were matched according to sex, age and BMI. Patients had a BMI of more than 40, or more than 35 with at least one other disease such as diabetes, but didn't exceed a BMI of 50.

The gastric-bypass patients had a higher rate of complications immediately after surgery. The study showed the early complication rate for gastric bypass was 17.2% compared with 5.4% for banding. But in the long term, there were more complications and more follow-up operations after gastric banding.

Robin Blackstone, the president of the American Society for Metabolic and Bariatric Surgery who practices in Scottsdale, Ariz., explained that gastric banding initially seemed safer than gastric-bypass surgery in the first three months after the procedure. She says the long-term data involving both procedures will help doctors better determine treatment choices for obese patients. "What's important is how effective both procedures are," she

ONLINE FIRST

Roux-en-Y Gastric Bypass vs Gastric Banding for Morbid Obesity

A Case-Matched Study of 442 Patients

Sébastien Romy, MD; Andrea Donadini, MD; Vittorio Giusti, MD, PD; Michel Suter, MD, Prof

Hypothesis: Gastric banding (GB) and Roux-en-Y gastric bypass (RYGBP) are used in the treatment of morbidly obese patients. We hypothesized that RYGBP provides superior results.

Design: Matched-pair study in patients with a body mass index (BMI) less than 50.

Setting: University hospital and regional community hospital with a common bariatric surgeon.

Patients: Four hundred forty-two patients were matched according to sex, age, and BMI.

Interventions: Laparoscopic GB or RYGBP.

Main Outcome Measures: Operative morbidity, weight loss, residual BMI, quality of life, food tolerance, lipid profile, and long-term morbidity.

Results: Follow-up was 92.3% at the end of the study period (6 years postoperatively). Early morbidity was higher after RYGBP than after GB (17.2% vs 5.4%; $P<.001$), but major morbidity was similar. Weight loss was quicker, maximal weight loss was greater, and weight loss remained significantly better after RYGBP until the sixth postoperative year. At 6 years, there were more failures (BMI > 35 or reversal of the procedure/conversion) after GB (48.3% vs 12.3%; $P<.001$). There were more long-term complications (41.6% vs 19%; $P<.001$) and more reoperations (26.7% vs 12.7%; $P<.001$) after GB. Comorbidities improved more after RYGBP.

Conclusions: Roux-en-Y gastric bypass is associated with better weight loss, resulting in a better correction of some comorbidities than GB, at the price of a higher early complication rate. This difference, however, is largely compensated by the much higher long-term complication and reoperation rates seen after GB.

Arch Surg. Published online January 16, 2012.
doi:10.1001/archsurg.2011.1708

THE PREVALENCE OF MORBID obesity has been growing exponentially over the past 20 years. A recent survey showed that bariatric procedures have more than doubled between 2003 and 2008.[1] In the United States, the increase was much greater for gastric banding (GB) than for gastric bypass (RYGBP).

See Invited Critique at end of article

This is probably because GB is perceived both by doctors and patients as a simple, safe, and reversible operation but also because of a huge industry-driven marketing campaign. Because GB was approved by the Food and Drug Administration only in 2001, the evolution in the United States is similar to that observed in Europe and Australia a decade before. In Europe, an opposite trend has recently been noted.[1]

Controversy about bariatric procedures has been ongoing. For patients with body mass index (BMI) less than 50 (calculated as weight in kilograms divided by height in meters squared), it lies mostly between purely restrictive operations (GB and vertical banded gastroplasty) and restrictive/malabsorptive procedures (RYGBP) also acting by hormone-mediated mechanisms influencing hunger and satiety.[2-4] Several trials have demonstrated the superiority of RYGBP over vertical banded gastroplasty regarding weight loss and long-term complications, resulting in the progressive abandonment of the latter.[5-10] Until now, 17 studies comparing GB with RYGBP have been published,[11-27] including 2 randomized trials,[11,12] 3 case-matched studies,[11,12,27] and many with important methodological flaws (eg, small numbers and different patient groups) and/or very limited follow-up. Tice et al[28] reviewed those available in 2008. General conclusions were that RYGBP pro-

Author Affiliations: Department of Visceral Surgery (Drs Romy, Donadini, and Suter) and Division of Endocrinology, Diabetology, and Metabolism (Dr Giusti), Centre Hospitalier Universitaire Vaudois, Lausanne, and Department of Surgery, Hôpital du Chablais, Aigle-Monthey (Dr Suter), Switzerland.

Class Example Stat222 (week4), Stat209 (week9) BK overview
urea synthesis, BK data data, long-form BK plots (by group)

2017 Analysis handout Extended BK lmer analysis (ascii)

3. Time 1 Time 2 observational data, Differences in Differences analysis.
We reuse some time-1, time-2 observational data generated to illustrate Lord's paradox (week 9, Stat209) -- gender differences in weight gain. (The 'paradox' is solved by Holland, Wainer, Rubin using potential outcomes.) The set up for these artificial data is females gain, males no change

```
  corr .7 within gender, equal vars time1 time 2 within gender
means
                 M              F
X (t1)         170            120
Y (t2)         170            130
```

comparison of "gains" 170 - 170 - (130 - 120) = -10 negative effect males (females gain more).
ancova: 170 - 130 - .7*(170 - 120) = 5 positive male effect
So: does being male cause a student to gain weight or lose weight? Illustrate forms of diffs-in-diffs analyses.
wide form for these data long form for these data
Solution for Review Question 3

Week 8-- Instrumental Variable Methods for Randomized Controlled Trials

In the news
Better Diet Tied to Bigger Brains Dutch study shows association between food and brain structure .
Publication: Croll, Pauline H. et al Better diet quality relates to larger brain tissue volumes. Neurology (2018): Web. 20 May. 2018.

Lecture Topics Lecture 8 slide deck

```
Encouragement design (Holland 1988 )
Instrumental variable methods for causal inference ( Baiocchi, Cheng and Small 2004)
```

Lecture 7 addendum: **Case-control studies**
Breslow NE. Statistics in epidemiology: the case-control study.J Am Stat Assoc. 1996 Mar;91(433):14-28
Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma JNCI: Journal of the National Cancer Institute, Volume 98, Issue 1, 4 January 2006, Pages 72-75,
Smoking and Lung Cancer in Chap 18 of HSAUR3 (Handbook of Statistical Analysis Using R). Also driving and backpain data in Chap 7 HSAUR2
Some R-packages and resources: SensitivityCaseControl: Sensitivity Analysis for Case-Control Studies; multipleNCC: Inverse Probability Weighting of Nested Case-Control Data; Two-phase designs in epidemiology (Thomas Lumley) ; Exact McNemar's Test and Matching Confidence Intervals

Computing Corner: **Regression Discontinuity Designs**
Example from rdd manual (Stat209 handout) ascii version
Angrist-Lavy Maimondes (class size) data sections 1.3, 3.2, 5.2.3, 5.3 DOS text
read data `ang = read.dta("http://stats.idre.ucla.edu/stat/stata/examples/methods_matter/chapter9/angrist.dta")`
R-package--rdd; Regression Discontinuity Estimation Author Drew Dimmery
Also Package `rdrobust` Title Robust data-driven statistical inference in Regression-Discontinuity designs
Slides for Regression Discontinuity CC
Regression Discontinuity Resources
Stat209, Regression Discontinuity handout
William Trochim's Knowledge Base
Trochim W.M. & Cappelleri J.C. (1992). "Cutoff assignment strategies for enhancing randomized clinical trials." Controlled Clinical Trials, 13, 190-212. pubmed link
Journal of Econometrics (special issue) Volume 142, Issue 2, February 2008, The regression discontinuity design: Theory and applications Regression discontinuity designs: A guide to practice, Guido W. Imbens, Thomas Lemieux

Statistics in Epidemiology: The Case-Control Study

N. E. BRESLOW

Statisticians have contributed enormously to the conceptualization, development, and success of case-control methods for the study of disease causation and prevention. This article reviews the major developments. It starts with Cornfield's demonstration of odds ratio invariance under cohort versus case-control sampling, proceeds through the still-popular Mantel–Haenszel procedure and its extensions for dependent data, and highlights (conditional) likelihood methods for relative risk regression. Recent work on nested case-control, case-cohort, and two-stage case-control designs demonstrates the continuing impact of statistical thinking on epidemiology. The influence of R. A. Fisher's work on these developments is mentioned wherever possible. His objections to the drawing of causal conclusions from observational data on cigarette smoking and lung cancer are used to introduce the problems of measurement error and confounding bias. The resolution of such difficulties, whether by further development and implementation of randomized intervention trials or by causal analysis of observational data using graphical models containing latent variables, will challenge future generations of statisticians.

KEY WORDS: Likelihood; Mantel–Haenszel procedure; Matched samples; Observational data; Odds ratio; Relative risk regression.

1. INTRODUCTION

The sophisticated use and understanding of case-control studies is the most outstanding methodologic development of modern epidemiology (Rothman 1986, p. 62).

My choice of topic for the 1995 Fisher Lecture is based on my belief that the contributions made by statisticians to the development of case-control methodology over the past 50 years have been among the most important of the many contributions they have made to public health and biomedicine. This view is shared by many epidemiologists. Writing in the first 1994 issue of *Epidemiologic Reviews,* which was devoted entirely to applications of the case-control method, Armenian and Lilienfeld (1994, p. 3) declared that the impact of statisticians on the "development of epidemiology would be difficult to overstate." Rothman's quotation, from his influential textbook *Modern Epidemiology,* highlights the importance of case-control methods in current epidemiologic research. The continuing popularity of the methodology is evident from the fact that 223 population-based case-control studies were published in the world literature in 1992 (Correa, Stewart, Yeh, and Santos-Burgoa 1994).

I am most grateful to the Committee and to the Organizers for the invitation to present the 1995 Fisher Lecture and for the opportunity to discuss a subject that has stimulated much of my research work. I would like to acknowledge Professors L. Moses and B. Efron, my graduate and dissertation advisors; Professor P. Armitage, who hosted me during a seminal postdoctoral year; and above all Professor N. Day, who introduced me to case-control studies and with whom I have enjoyed a long and fruitful collaboration. It is also a pleasure to acknowledge the outstanding contributions made to this field, and to my understanding of it, by my colleagues and by a score of graduates of the University of Washington Biostatistics Program.

2. ORIGINS

The central idea of the case-control study is the comparison of a group having the outcome of interest to a control group with regard to one or more characteristics. An early example is Guy's 1843 comparison of the occupations of men with pulmonary consumption to those of men with other diseases (Lilienfeld and Lilienfeld 1979). The method became popular during the 1920s for the study of cancer, notable successes being the associations discovered between lip cancer and pipe smoking by Broders (1920), between breast cancer and reproductive history by Lane-Claypon (1926), and between oral cancer and pipe smoking by Lombard and Doering (1928). Because these diseases were rare, it was rather impractical to study them in any other way; for example, by follow-up of an initially healthy population. Increased attention to *and criticism of* case-control methodology followed the publication in 1950 of several studies of smoking and lung cancer (Surgeon General 1964).

Under the leadership of Harold Dorn, statisticians at the U.S. National Cancer Institute were stimulated by the ensuing controversy to investigate the advantages and shortcomings of the case-control method. A prevailing belief at the time was that separate samples of cases and controls did not provide relevant quantitative information about the parameters of primary interest—namely, the disease rates. This misconception was corrected by Jerome Cornfield (1951), who is widely credited with launching the modern era of case-control studies. Cornfield demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed, and that the latter in turn approximates the ratio of disease rates provided that the disease is rare. Formally, if D denotes disease (1 for cases, 0 for controls) and X denotes exposure (1 for

N. E. Breslow is Professor of Biostatistics, University of Washington, Seattle, WA, 98195. This article is based on the R. A. Fisher Lecture delivered to the Joint Statistical Meetings, Orlando, FL 1995. The work was supported in part by U.S. Public Health Service Grant CA40644.

case-control

Case-Control Study

The case-control study, a widely used method of observational epidemiological study, is an application of medical history-taking that aims to identify the cause of disease among a group of people, or the cause-effect relationships of a condition of interest. The underlying concept is simple. The past medical history, or history of exposure to a suspected risk or protective factor, of a group of persons with the disease or condition of interest (the cases) is compared with the past history of another group of persons (the controls) who resemble them in as many relevant respects as possible, but who do not have the disease or condition of interest. Statistical analysis is used to determine whether there is a stronger association of past exposure to the suspected risk or protective factor with the condition of interest among the cases than among the controls. The method can be called a retrospective study because it is concerned with events in the past. However, the cases are often collected prospectively, with cases added as they occur, so there is possible confusion with what used to be called a prospective study but is now almost always called a cohort study. It has also been called case-compeer study and case-referent study, but case-control study is the most widely used term.

The method evolved out of analyses of series of cases. The concept was mentioned in the writings of the nineteenth-century French physician Pierre Charles Alexandre Louis and a simple form of it was used by the nineteenth-century English physician William Augustus Guy. In the 1930s, the English physician Janet Lane-Claypon used this method to study risk factors for breast cancer, and in 1939, just as war was breaking out in Europe, F. H. Muller, a German physician, used a case-control study design to demonstrate that a past history of cigarette smoking was strongly associated with lung cancer. Following World War II, several investigators in England and in the United States adopted Muller's methods for case-control studies of smoking and lung cancer, which had become a very common and lethal form of cancer. In 1950, Doll and Hill in the England and Wynder and Graham in the United States published large case-control studies of cigarette smoking and cancer of the lung almost simultaneously in the *British Medical Journal* and the *Journal of the American Medical Association*, respectively. Many more case-control studies of this and other kinds of cancer soon established the utility of the method.

Case-control studies have proved particularly useful in studying very rare conditions. During 1969 and 1970, eight case of adenocarcinoma of the vagina were seen in adolescent girls and young women in Boston, Massachusetts. This was, up till then, an extremely rare, almost nonexistent condition, and it was clear that these young women must have been exposed to some unusual cancer-causing agent. Each of the eight cases was matched with four otherwise similar but healthy females of the same age. Their, and their mothers', past histories of many kinds of exposure to medications, vaginal douches, and other substances, were compared. Seven of the eight cases had a history of their mothers having been given artificial estrogen to prevent miscarriage early in pregnancy (this had been a popular though unproven method of preventing threatened miscarriage since the 1950s; it has now been shown to be useless). None of the controls had a similar history. There was less than a 1 in 100,000 likelihood of this distribution occurring by chance. Adenocarcinoma of the vagina was caused by prenatal exposure of the developing female fetus to diethylstilbestrol, an artificial estrogen. Later studies showed that genital dysplasia in boys and young men was another consequence of prenatal exposure to artificial estrogen.

These examples, and many others, illustrate the value of the case-control study. It is a relatively cheap, rapid, and reliable method of establishing evidence of an association between an exposure to a risk (or protective) factor and an unfavorable (or favorable) outcome. It does not require study of large numbers. The concept is readily understandable, so members of the lay public, political decision makers, and the media can easily grasp the significance of the findings.

There are, however, some important shortcomings. The results can be biased in many waysâ€" by flawed information about past exposure to risk, inappropriate selection of controls, and various confounding factors. The validity of results based on the use of controls who may have been exposed to similar or different combinations of risk, biases introduced by selective recall or recording of relevant past exposure to risk, and the most suitable way to analyze the results have generated endless debates in epidemiological journals.

The advantages of the case-control method are: (1) it is an excellent way to study rare diseases and diseases with long latency, (2) a relatively quick answer can be obtained, (3) it is relatively cheap, (4) it usually requires only a few cases, (5) it can often make use of existing records, and (6) it can study several possible causes or exposures to risk simultaneously.

The disadvantages of the method are: (1) it relies on subjects' recall and/or completeness of existing records, (2) it may be difficult or impossible to validate this information, (3) there is incomplete allowance for extraneous factors, (4) the selection of a suitable comparison (control) group may be difficult, (5) rates cannot be calculated, (6) the mechanism of disease cannot be studied, and(7) a proof of causation cannot be established.

ex from last year

JNCI Journal of the National Cancer Institute

Journal of the National Cancer Institute, Vol. 98, No. 1, 72-75, January 4, 2006
DOI: 10.1093/jnci/djj007

not an experiment, randomized controlled field trial, case-control (see SW, instead) and there couched "it is not likely"

This Article
Full Text
Full Text (PDF)
Alert me when this article is cited
Alert me if a correction is posted
Services
Email this article to a friend
Alert me to new issues of the journal
Download to citation manager
Cited by other HighWire-hosted articles
Search for citing articles in: ISI Web of Science (3)
Request Permissions
Disclaimer
Google Scholar
Articles by Mayne, S. T.
Articles by Fraumeni, J. F.
Articles citing this Article
PubMed
PubMed Citation
Articles by Mayne, S. T.
Articles by Fraumeni, J. F., Jr.
Related Collections
Get Related:
-JNCI Articles
-Journal Articles
-Cancer Statistics
-PDQ Summaries/Trials
-PDR-Physicians' Desk Reference
-Cochrane Reviews
-Useful Links
Correspondence about this Article

BRIEF COMMUNICATION

Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma

Susan T. Mayne, Harvey A. Risch, Robert Dubrow, Wong-Ho Chow, Marilie D. Gammon, Thomas L. Vaughan, Lauren Borchardt, Janet B. Schoenberg, Janet L. Stanford, A. Brian West, Heidi Rotterdam, William J. Blot, Joseph F. Fraumeni, Jr.

Affiliations of authors: Yale University School of Medicine and Yale Cancer Center, Department of Epidemiology and Public Health, New Haven, CT (STM, HAR, RD, LB); National Cancer Institute, Division of Cancer Epidemiology and Genetics, NIH, DHHS, Bethesda, MD (W-HC, JFF); University of North Carolina, School of Public Health, Department of Epidemiology, Chapel Hill, NC (MDG); Fred Hutchinson Cancer Research Center, Program in Epidemiology, and University of Washington, School of Public Health and Community Medicine, Department of Epidemiology, Seattle, WA (TLV, JLS); New Jersey Department of Health and Senior Services, Center for Cancer Initiatives, Trenton, NJ (JBS); Columbia University, Department of Pathology, New York, NY (HR); New York University Medical Center, Department of Pathology, New York, NY (ABW); International Epidemiology Institute, Rockville, MD (WJB)

Correspondence to: Susan T. Mayne, PhD, Yale University School of Medicine, Department of Epidemiology and Public Health, 60 College Street, New Haven, CT 06520-8034 (e-mail: Susan.Mayne{at}yale.edu).

Carbonated soft drinks (CSDs) have been associated with gastroesophageal reflux, an established risk factor for esophageal adenocarcinoma. As both CSD consumption and esophageal adenocarcinoma incidence have sharply increased in recent decades, we examined CSD as a risk factor for esophageal and gastric cancers in a U.S. multicenter, population-based case-control study. Associations between CSD intake and risk were estimated by adjusted odds ratios (ORs), comparing the highest versus lowest quartiles of intake. All statistical tests were two-sided. Contrary to the proposed hypothesis, CSD consumption was inversely associated with esophageal adenocarcinoma risk (highest versus lowest quartiles, OR = 0.47, 95% confidence interval = 0.29 to 0.76; P_{trend} = .005), due primarily to intake of diet CSD. High CSD consumption did not increase risk of any esophageal or gastric cancer subtype in men or women or when analyses were restricted to nonproxy interviews. These findings indicate that CSD consumption (especially diet CSD) is inversely associated with risk of esophageal adenocarcinoma, and thus it is not likely to have contributed to the rising incidence rates.

This article has been cited by other articles in HighWire Press-hosted journals:

- Lim, U., Subar, A. F., Mouw, T., Hartge, P., Morton, L. M., Stolzenberg-Solomon, R., Campbell, D., Hollenbeck, A. R., Schatzkin, A. (2006). Consumption of Aspartame-Containing Beverages and Incidence of Hematopoietic and Brain Malignancies. *Cancer Epidemiol Biomarkers Prev* 15: 1654-1659 [Abstract] [Full Text]
- Lagergren, J., Viklund, P., Jansson, C. (2006). Carbonated Soft Drinks and Risk of Esophageal Adenocarcinoma: A Population-Based Case-Control Study. *J Natl Cancer Inst* 98: 1158-1161 [Abstract] [Full Text]
- Gallus, S., Talamini, R., Fernandez, E., Dal Maso, L., Franceschi, S., La Vecchia, C. (2006). Re: Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma. *J Natl Cancer Inst* 98: 645-646 [Full Text]
- Mallath, M. K. (2006). Re: Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma. *J Natl Cancer Inst* 98: 644-645 [Full Text]

Correspondence about this Article

- CORRESPONDENCE
  Mohandas K. Mallath
  Re: Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma
  J Natl Cancer Inst 2006; 98: 644-645. [Extract] [Full Text] [PDF]

- CORRESPONDENCE
  Silvano Gallus, Renato Talamini, Esteve Fernandez, Luigino Dal Maso, Silvia Franceschi, and Carlo La Vecchia
  Re: Carbonated Soft Drink Consumption and Risk of Esophageal Adenocarcinoma
  J Natl Cancer Inst 2006; 98: 645-646. [Extract] [Full Text] [PDF]

association background: esoph canc and soda both increase over years

Rosenbaum text: case-control

DES and Vaginal Cancer: Sensitivity to Bias

Cancer of the vagina is a rare condition, particularly in young women. In 1971, Herbst, Ulfelder, and Poskanzer published a report describing eight cases of vaginal cancer in women aged 15 to 22. They were particularly interested in the possibility that a drug, diethylstilbestrol or DES, given to pregnant women, might be a cause of vaginal cancer in their daughters. Each of the eight cases was matched to four *referents*, that is, to four women who did not develop vaginal cancer. These four referents were born within five days of the birth of the case at the same hospital, and on the same type of service, ward or private. There were then eight cases of vaginal cancer and 32 referents, and the study compared the use of DES by their mothers.

This sort of study is called ~~a *case-referent study* or a~~ *case-control study* or a *retrospective study*, no one terminology being universally accepted. In an experiment and in many observational studies, treated and control groups are followed forward in time to see how outcomes develop. In the current context, this would mean comparing two groups of women, a treated group whose mothers had received DES and a control group whose mothers had not. That sort of study is not practical because the outcome, vaginal cancer, is so rare—the treated and control groups would have to be enormous and continue for many years to yield eight cases of vaginal cancer. In a case-referent study, the groups compared are not defined by whether or not they received the treatment, but rather by whether or not they exhibit the outcome. The cases are compared to the referents to see if exposure to the treatment is more common among cases.

In general, the name "case-control" study is not ideal because the word "control" does not have its usual meaning of a person who did not receive the treatment. In fact, in most case-referent studies, many referents did receive the treatment. The name "retrospective" study is not ideal because there are observational studies in which data on entire treated and control groups are collected after treatments have been given and outcomes have appeared, that is, collected retrospectively, and yet the groups being compared are still treated and untreated groups. See MacMahon and Pugh (1970, pp. 41–46) for some detailed discussion of this terminology.

So the study compared eight cases of vaginal cancer to 32 matched referents to see if treatment with diethylstilbestrol was more common among mothers of the cases, and indeed it was. Among the mothers of the eight cases, seven had received DES during pregnancy. Among mothers of the 32 referents, none had received DES. The association between vaginal cancer and DES appears to be almost as strong as a relationship can be, though of course only eight cases have been observed. If a conventional test designed for use in a randomized experiment is used to compare cases and referents in terms of the frequency of exposure to DES, the difference is highly significant. However, experience with the first example, vitamin C and cancer, suggests caution here.

4.4.5 *Matched Case-Referent Studies: Five Examples*

Recall from Chapter 1 that in the case-referent study by Herbst, Ulfelder, and Poskanzer (1971) of DES and vaginal cancer, each of eight cases of vaginal cancer was matched to four referents using day of birth and type of service, ward or private. Table 4.8 contains the data together with some of the calculations for the sensitivity analysis.

In Table 4.8, there are $S = 8$ matched sets, $s = 1, \ldots, 8$, each containing one case and four referents. In seven matched sets, exactly one patient had in utero exposure to DES, that is, $m_s = 1$, but in set $s = 5$ there were no exposures to DES, $m_s = 0$. In all of the seven matched sets, it was the case who was exposed to DES, that is, $B_s = 1$ for $s \neq 5$. The relationship between DES and vaginal cancer appears to be extremely strong, though

TABLE 4.8. DES and Vaginal Cancer: Data and Computations.

s	B_s	m_s	$P_s^- = P_s^+$ for $\Gamma = 1$	P_s^- for $\Gamma = 2$	P_s^+ for $\Gamma = 2$
1	1	1	0.20	0.11	0.33
2	1	1	0.20	0.11	0.33
3	1	1	0.20	0.11	0.33
4	1	1	0.20	0.11	0.33
5	0	0	0.00	0.00	0.00
6	1	1	0.20	0.11	0.33
7	1	1	0.20	0.11	0.33
8	1	1	0.20	0.11	0.33
Total	7	7	1.40	0.78	2.33

of course there are only eight cases. The one matched set with no exposed patients is concordant, and it could be removed without changing the value of the statistics in (4.20).

To test the null hypothesis of no effect, suppose for the moment that DES does not cause vaginal cancer. Consider one of the seven matched sets with $n_s = 5$ patients of whom $m_s = 1$ was exposed to DES. If the study were free of hidden bias, so $\Gamma = 1$, then each of the five patients has the same chance of being exposed, so the chance that the one exposed patient is the case is 1/5 or 0.20, as in Table 4.8. This would lead us to expect $1.40 = 7 \times 0.20$ cases to be exposed to DES, though, in fact, $T = 7$ cases were exposed, with a variance of $7 \times 0.20 \times 0.80 = 1.12$, and a deviate of $(|7 - 1.4| - \frac{1}{2})/\sqrt{1.12} = 4.82$, whose square $4.82^2 = 23.2$ is the usual Mantel–Haenszel statistic. If the study were free of hidden bias, there would be strong evidence that DES causes vaginal cancer. If hidden bias were present to the extent that matched subjects might differ in their odds of exposure to DES by a factor of two, so $\Gamma = 2$, then the chance that the case was exposed to DES might be as low as $m_s/\{m_s + \Gamma(n_s - m_s)\} = 1/(1+2\times4) = 0.11$ or as high as $\Gamma m_s/\{\Gamma m_s + (n_s - m_s)\} = 2/(2+4) = 0.33$, so the expected number of exposed cases might be as low as $7 \times 0.11 = 0.78$ with variance $7 \times 0.11 \times 0.89 = 0.69$, or as high as $7 \times 0.33 = 2.33$ with variance $7 \times 0.33 \times 0.67 = 1.56$, whereas seven exposed cases were observed. For $\Gamma = 2$, the deviates in (4.20) are

$$\frac{(|7 - 2.33| - \frac{1}{2})}{\sqrt{1.56}} = 3.34 \quad \text{and} \quad \frac{(|7 - 0.78| - \frac{1}{2})}{\sqrt{0.69}} = 6.88,$$

yielding a range of significance levels from less than 0.0001 to at most 0.0004. A hidden bias of magnitude $\Gamma = 2$ cannot reasonably explain the strong association seen between DES and vaginal cancer. Table 4.9 gives results for other values of Γ. Only beyond $\Gamma = 7$ is hidden bias a plausible

A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 18

Incorporating Prior Knowledge via Bayesian Inference: Smoking and Lung Cancer

18.1 Introduction

At the beginning of the 20th century, the death toll due to lung cancer was on the rise and the search for possible causes began. For lung cancer in pit workers, animal experiments showed that the so-called 'Schneeberg lung disease' was induced by radiation. But this could not explain the increasing incidence of lung cancer in the general population. The identification of possible risk factors was a challenge for epidemiology and statistics, both disciplines being still in their infancy in the 1920s and 1930s.

The first modern controlled epidemiological study on the effect of smoking on lung cancer was performed by Franz Hermann Müller as part of his dissertation at the University of Cologne in 1939. The results were published a year later (**?**). Müller sent out questionnaires to the relatives of people who had recently died of lung cancer, asking about the smoking behavior and its intensity of the deceased relative. He also sent the questionnaire to healthy controls to obtain information about the smoking behavior in a control group, although it is not clear how this control group was defined. The number of lung cancer patients and healthy controls in five different groups (nonsmokers to extreme smokers) are given in Table 18.1.

Table 18.1: `Smoking_Mueller1940` data. Smoking and lung cancer case-control study by Müller (1940). The smoking intensities were defined by the number of cigarettes smoked daily: 1-15 (moderate), 16-25 (heavy), 26-35 (very heavy), and more than 35 (extreme).

		`Diagnosis`	
`Smoking`		Lung cancer	Healthy control
	Nonsmoker	3	14
	Moderate smoker	27	41
	Heavy smoker	13	22
	Very heavy smoker	18	5
	Extreme smoker	25	4

Four years later Erich Schöninger also wrote his dissertation on the association between smoking and lung cancer and, together with his supervisor Eberhard Schairer at the University of Jena, published his results on a case-control study (**?**) where he assessed the smoking behavior of lung cancer patients, patients diagnosed with other forms of cancer, and also a healthy control group. The data are given in Table 18.2.

Table 18.2: `Smoking_SchairerSchoeniger1944` data. Smoking and lung cancer case-control study by Schairer and Schöniger (1944). Cancer other than lung cancer omitted. The smoking intensities were defined by the number of cigarettes smoked daily: 1-5 (moderate), 6-10 (medium), 11-20 (heavy), and more than 20 (very heavy).

	Diagnosis	
`Smoking`	Lung cancer	Healthy control
Nonsmoker	3	43
Moderate smoker	11	98
Medium smoker	31	57
Heavy smoker	19	47
Very heavy smoker	29	25

Shortly after the war, a Dutch epidemiologist reported on a case-control study performed in Amsterdam (**?**) and found similar results as the two German studies; see Table 18.3.

Table 18.3: `Smoking_Wassink1945` data. Smoking and lung cancer case-control study by Wassink (1945). Smoking categories correspond to the categories used by Müller (1940).

	Diagnosis	
`Smoking`	Lung cancer	Healthy control
Nonsmoker	6	19
Moderate smoker	18	36
Heavy smoker	36	25
Very heavy smoker	74	20

In 1950 perhaps the most important, but not the first, case-control study showing an increasing risk of developing lung cancer with the amount of tobacco smoked, was published in Great Britain by Richard Doll and Austin Bradford Hill (**?**). We restrict discussion here to data obtained for males and the

data shown in Table 18.4 corresponds to the most recent amount of tobacco consumed regularly by smokers before disease onset (Table V in **?**).

Table 18.4: `Smoking_DollHill1950` data. Smoking and lung cancer case-control study (only males) by Doll and Hill (1950). The labels for the smoking categories give the number of cigarettes smoked every day.

		Diagnosis	
Smoking		Lung cancer	Other
	Nonsmoker	2	27
	1-	33	55
	5-	250	293
	15-	196	190
	25-	136	71
	50+	32	13

Although the design of the studies by **?** and **?**, especially the selection of their control groups, can be criticized (see **?**, for a detailed discussion) and the study by **?** was larger than the older studies and more detailed information on the smoking behavior was obtained by direct patient interviews, the information provided by the earlier studies was not taken into account by **?**. They cite **?** in their introduction, but did not compare their findings with his results. It is remarkable to see that both **?** and **?** extensively made use of the report by **?** and go as far as analyzing the merged data (Grafiek I, E, and F, in **?**). In an informal way, these authors wanted to use the already available information, in today's terms called 'prior knowledge', to make a stronger case with the new data. Formal statistical methods to incorporate prior knowledge into data analysis as part of the 'Bayesian' way of doing statistical analyses were developed in the second half of the last century, and we will focus on them in the present chapter.

18.2 Bayesian Inference

18.3 Analysis Using R

18.3.1 One-by-one Analysis

For the analysis of the four different case-control studies on smoking and lung cancer, we will (retrospectively, of course) update our knowledge with every new study. We begin with a re-analysis of the data described by **?**. Using an approximate permutation test introduced in Chapter **??** for the hypothesis of independence of the amount of tobacco smoked and group membership (lung cancer or healthy control), we get

matching to reduce bias in observational studies

Y ~ G , fix G nonequivalence

Modeling Selection Effects

Draft

21 January 2005

by Thad Dunning, Political Science Department
and David Freedman, Statistics Department
UC Berkeley, CA 94720

1. Introduction

Selection bias is a pervasive issue in social science. Three research topics illustrate the point.

(i) What are the returns to education? College graduates earn more than high school graduates, but the difference could be due to factors—like intelligence and family background—that lead some persons to get a college degree while others stop after high school.

(ii) Are job training programs effective? If people who take the training are relatively ambitious and well organized, any direct comparison is likely to over-estimate program effectiveness, because participants are more likely to find employment anyway. (See references below.)

(iii) Do boot camps for prisoners prevent recidivism? Possibly, but prisoners who want to go straight are more likely to participate, and less likely to find themselves in jail again—even if boot camp has no effect.

These questions could be settled by experiment, but experimentation in such contexts is expensive at best, impractical or unethical at worst. Investigators rely, therefore, on observational (non-experimental) data, with attendant difficulties of confounding.

In brief, comparisons can be made between a treatment group and a control group that does not get the treatment. But there are likely to be differences between the groups, other than the treatment. Such differences are called "confounding factors." Differences on the response variable of interest (income, employment, recidivism) may be due to treatment, or confounding factors, or both. Confounding is especially troublesome when subjects select themselves into one group or another, rather than being assigned to different regimes by the investigator. Self-selection is the hallmark of an observational study; assignment by the investigator is the hallmark of an experiment.

This article will review one of the most popular models for selection bias. The model, due to Heckman, will be illustrated on the relationship between admissions tests and college grades. Causal inference will be mentioned. There will be some pointers to the literature on selection bias, including critiques and alternative models. The intention-to-treat principle for clinical trials will be discussed, by way of counterpoint.

Model-based corrections for selection bias turn out to depend strongly on the assumptions built into the model. Thus, caution is in order. Sensitivity analysis is highly recommended: try different models with different assumptions. Alternative research designs should also be considered: stronger designs may permit data analysis with weaker assumptions.

2. Admissions data

In the US, many colleges and universities require applicants to take the SAT (Scholastic Achievement Test). Admission is based in part on SAT scores and in part on other evidence—high school GPA (grade point average), essays, recommendations, interviews by admissions officers. Figure 1 shows a somewhat hypothetical scatter diagram. Each student is represented by a dot. The

Sample Selection Models in R: Package sampleSelection

Ott Toomet
Tartu University

Arne Henningsen
University of Copenhagen

Abstract

This introduction to the R package **sampleSelection** is a slightly modified version of Toomet and Henningsen (2008b), published in the *Journal of Statistical Software*.

This paper describes the implementation of Heckman-type sample selection models in R. We discuss the sample selection problem as well as the Heckman solution to it, and argue that although modern econometrics has non- and semiparametric estimation methods in its toolbox, Heckman models are an integral part of the modern applied analysis and econometrics syllabus. We describe the implementation of these models in the package **sampleSelection** and illustrate the usage of the package on several simulation and real data examples. Our examples demonstrate the effect of exclusion restrictions, identification at infinity and misspecification. We argue that the package can be used both in applied research and teaching.

Keywords: sample selection models, Heckman selection models, econometrics, R.

1. Introduction

Social scientists are often interested in causal effects—what is the impact of a new drug, a certain type of school or being born as a twin. Many of these cases are not under the researcher's control. Often, the subjects can decide themselves, whether they take a drug or which school they attend. They cannot control whether they are twins, but neither can the researcher—the twins may tend to be born in different types of families than singles. All these cases are similar from the statistical point of view. Whatever is the sampling mechanism, from an initial "random" sample we extract a sample of interest, which may not be representative of the population as a whole (see Heckman and MaCurdy 1986, p. 1937, for a discussion).

This problem—people who are "treated" may be different than the rest of the population—is usually referred to as a *sample selection* or *self-selection* problem. We cannot estimate the causal effect, unless we solve the selection problem[1]. Otherwise, we will never know which part of the observable outcome is related to the causal relationship and which part is due to the fact that different people were selected for the treatment and control groups.

Solving sample selection problems requires additional information. This information may be in different forms, each of which may or may not be feasible or useful for any particular case.

[1]Correcting for selectivity is necessary but not always sufficient for estimating the causal effect. Another common problem is the lack of common support between the treated and untreated population. We are grateful to a referee for pointing this out.

Here we list a few popular choices:

- Random experiment, the situation where the participants do not have control over their status but the researcher does. Randomisation is often the best possible method as it is easy to analyse and understand. However, this method is seldom feasible for practical and ethical reasons. Even more, the experimental environment may add additional interference which complicates the analysis.
- Instruments (exclusion restrictions) are in many ways similar to randomisation. These are variables, observable to the researcher, and which determine the treatment status but not the outcome. Unfortunately, these two requirements tend to contradict each other, and only rarely do we have instruments of reasonable quality.
- Information about the functional form of the selection and outcome processes, such as the distribution of the disturbance terms. The original Heckman's solution belongs to this group. However, the functional form assumptions are usually hard to justify.

During recent decades, either randomisation or pseudo-randomisation (natural experiments) have become state of the art for estimating causal effects. However, methods relying on distributional assumptions are still widely used. The reason is obvious—these methods are simple, widely available in software packages, and they are part of the common econometrics syllabus. This is true even though reasonable instruments and parametric assumptions can only seldom be justified, and therefore, it may be hard to disentangle real causal effects from (artificial) effects of parametric assumptions.

Heckman-type selection models also serve as excellent teaching tools. They are extensively explained in many recent econometric text books (e.g. Johnston and DiNardo 1997; Verbeek 2000; Greene 2002; Wooldridge 2003; Cameron and Trivedi 2005) and they are standard procedures in popular software packages like **Limdep** (Greene 2007) and Stata (StataCorp. 2007). These models easily allow us to experiment with selection bias, misspecification, exclusion restrictions etc. They are easy to implement, to visualize, and to understand.

The aim of this paper is to describe the R (R Development Core Team 2008) package **sampleSelection** (version 0.6-0), which is available on the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=sampleSelection`. The package implements two types of more popular Heckman selection models which, as far as we know, were not available for R before. Our presentation is geared toward teaching because we believe that one of the advantages of these types of models lies in econometrics training.

The paper is organized as follows: In the next section we introduce the Heckman (1976) solution to the sample selection problem. Section 3 briefly describes the current implementation of the model in **sampleSelection** and its possible generalisations. In Section 4 we illustrate the usage of the package on various simulated data sets. Section 5 is devoted to replication exercises where we compare our results to examples in the literature. Section 6 describes robustness issues of the method and our implementation of it; and the last section concludes.

2. Heckman's solution

The most popular solutions for sample selection problems are based on Heckman (1976). A variety of generalisations of Heckman's standard sample selection model can be found in the

STANFORD UNIVERSITY

obs studies checklist

1) Limit outcome comparison to regions (persons) where you have data [interp ~~extrap~~] (common support / subsamples)

2) Use X's for the purpose they are intended (correct, redress imbalance from self-selection etc)

3) Use X's in a robust manner (not ~~sensitive~~ sensitive (much) to what's included, measurement, funct form) i.e. fit rather than coeff.

4) Be able to see how well the X's function to restore balance (and remediate)

Don't use Y / reserve outcomes

for as if by experiment statement

recreate the hypothetical experiment

1.5
1.38

Design trumps analysis, week 2

3.7. *The result.* These six steps combine to make for objective observational study design in the sense that the resultant designed study can be conceptualized as a hypothetical, approximating randomized block (or paired comparison) experiment, whose blocks (or matched pairs) are our balancing groups, and where the probabilities of treatment versus control assignment may vary relatively dramatically across the blocks. This statement does not mean the researcher who follows these steps will achieve an answer similar to the one that would have been found in the analogous randomized experiment, but at least the observational study has a chance of doing so, whereas if these steps are not followed, I believe that it is only blind luck that could lead to a similar answer as in the analogous randomized experiment.

Sometimes the design effort can be so extensive that a description of it, with no analyses of any outcome data, can be itself publishable. For a specific example on peer influence on smoking behaviors, see Langenskold and Rubin (2008).

4. Examples using propensity scores and subclassification.

4.1. *Classic example with one observed covariate.* The following very simple example is taken from Cochran (1968) classic article on subclassification in observational studies, which uses some smoking data to illustrate ideas. Let us suppose that we want to compare death rates (the outcome variable of primary interest) among smoking males in the U.S., where the treatment condition is considered cigarette smoking and the control condition is cigar and pipe smoking. There exists a very large dataset with the death rates of smoking males in the U.S., and it distinguishes between these two types of smokers. So far, so good, in that we have a dataset with Y and treatment indicators, and it is very large. Now we strip this dataset of all outcome data; no survival (i.e., Y) data are left and are held out of sight until the design phase is complete.

Next we ask (in a simple minded way, because this is only an illustrative example), who is the decision maker for treatment versus control, and what are the key covariates used to make this decision? It is relatively obvious that the main decision maker is the individual male smoker. It is also relatively obvious that the dominant covariate used to make this decision is age—most smokers start in their teens, and most start by smoking cigarettes, not pipes or cigars. Some pipe and cigar smokers start in college, but many start later in life. Cigarette smokers tend to have a more uniform distribution of ages. Other possible candidate key covariates are education, socio-economic status, occupational status, income, and so forth, all of which tend to be correlated with age, so to illustrate, we focus on age as our only X variable. Then our hypothetical randomized experiment starts with male smokers and randomly assigns them to cigarette or cigar/pipe smoking, where the propensity to be a cigarette smoker rather than a cigar/pipe smoker is viewed as a function of age. In this dataset, age is very well-measured. When we compare

the age distribution of cigarette smokers and age distribution of cigar/pipe smokers in the U.S. in this dataset, we see that the former are younger, but that there is substantial overlap in the distributions. Before moving on to the next step, we should worry about how people in the hypothetical experiment who died prior to the assembling of the observational dataset are represented, but, for simplicity in this illustrative example, we will move on to the next step.

How do we create subgroups of treatment and control males with more similar distributions of age than is seen overall, in fact, so similar that we could believe that the data arose from a randomized block experiment? Cochran's example used subclassification. First, the smokers are divided at the overall median into young smokers and old smokers—two subclasses, and then divided into young, middle aged, and old smokers, each of these three subclasses being equal size, and so forth. Finally, nine subclasses are used. The age distributions within each of the nine subclasses are very similar for the treatment condition and the control condition, just as if the men had been randomly assigned within the age subclasses to treatment and control, because there is such a narrow range of ages within each of the nine subclasses. And of great importance, there do exist both treatment and control males in each of nine subclasses.

The design phase can be considered complete for our simple illustrative example. Our underlying hypothetical randomized experiment that led to the observed dataset is a randomized block experiment with nine blocks defined by age, where the probability of being assigned to the treatment condition (cigarette smoking) rather than the control condition (cigar/pipe smoking) decreases with age. We are now allowed to look at the outcome data within each subclass and compare treatment and control death rates. We find that, averaging over the nine blocks (subclasses), the death rates are about 50% greater for the cigarette smokers than the cigar and pipe smokers. Incidentally, the full data set with no subclassification leads to nearly the opposite conclusion; see Cochran (1968) or Rubin (1997) for details.

But what would have happened if we decided that we wanted to subclassify also on education, socio-economic status, and income, each covariate using, let's say, five levels [a minimum number implicitly recommended in Cochran (1968)]? There would be four key covariates, each with five levels, yielding a total of 625 subclasses. And many observational studies have many more than four key covariates that are known to be used for making treatment decisions. For example, with 20 such covariates, even if each is dichotomous, there are 2^{20} subclasses—greater than a million, and as a result, many subclasses would probably have only one unit, either a treated or control, with no treatment-control comparison possible. How should we design this step of observational studies in such more realistic situations?

4.2. *Propensity score methodology.* Rosenbaum and Rubin (1983) proposed a class of methods to try to achieve balance in observational studies when there are many key covariates present. In recent years there has been an explosion of work

Subclassification to Balance Age

- To achieve balance on age, compare:
  - "young" cigar/pipe smokers with "young" cigarette smokers
  - "old" cigar/pipe smokers with "old" cigarette smokers
- Or better, compare:
  - Young, middle aged, old
  - Even more age subclasses
- Design phase, no outcome data, objective:
  - Approximates a randomized trial within subclasses
- Now look at outcome data

Nonrandomized Comparative Clinical Studies -
Proceedings of the International Conference on Nonrandomized
Comparative Clinical Studies in Heidelberg, April 10 -11,1997
Printed volume

Estimation from nonrandomized treatment comparisons using subclassification on propensity scores

(This article is a modification and expansion of an article in *Annals of Internal Medicine, 127*, 8(2), pp. 757-763.

D. B. Rubin

Abstract The aim of many analyses of medical data sets is to draw causal inferences about the relative effects of treatments, such as different methods of treating cancer patients. The data available to compare many such treatments are not based on the results of carefully conducted randomized clinical trials, but rather are collected while observing systems as they operate in "normal" practice, without any interventions implemented by randomized assignment rules. Such data are relatively inexpensive to obtain, however, and often do represent the spectrum of medical practice better than the settings of randomized experiments. Consequently, it is sensible to try to estimate the effects of treatments from such data sets, even if only to help design a new randomized experiment or shed light on the generalizability of results from existing randomized experiments. Standard methods of analysis using routine statistical software (e.g., linear or logistic regressions), however, can be quite deceptive for these objectives because they provide no warnings about their propriety. Propensity score methods are more reliable tools for addressing such objectives because the assumptions needed to make their answers appropriate are more assessable and transparent to the investigator. Subclassification on propensity scores is a particularly straightforward technique and is the topic of this article.

Propensity score technology in observational studies

The objective of many medical studies is the estimation of the causal effects of some new treatment or exposure relative to a control condition (e.g., the effect of smoking on mortality). In the vast majority of such studies, there is the need to control for naturally occurring systematic differences in background characteristics between the treatment group and the control group (e.g., in age or sex distributions), systematic differences which would not occur in the context of a randomized experiment. Typically, there are many background characteristics that need to be controlled.

Propensity score technology, introduced by Rosenbaum and Rubin (1983a), addresses this situation by reducing the entire collection of background characteristics to a single "composite" characteristic that appropriately summarizes the collection. This reduction from many characteristics to one composite characteristic allows the straightforward assessment of whether the treated and control groups overlap enough on background characteristics to allow sensible estimation of treatment versus control effects from this data set. Moreover, when such overlap is present, the propensity score approach allows straightforward calculation of estimated treatment versus control effects that reflect adjustment for differences in all observed background characteristics. Subclassification on the propensity score is a particularly straightforward technique for such adjustment.

Subclassification on One Confounding Variable

Before describing how subclassification on propensity scores can be used in the statistical analysis of an observational study with many confounding background characteristics, we begin with an example showing how subclassification can be used to adjust for a single confounding covariate, such as age, in a study of smoking and mortality. We then show how propensity scores methods can be used to generalize subclassification on a single confounding covariate to the case with many confounding covariates, such as age, region of the country, and sex. The potential for an observational data base (i.e., not from a randomized experiment) to suggest causal effects of treatments is indicated by Table 1, adapted from Cochran (1968), which concerns mortality rates per thousand in three large data bases from the U.S., the U.K., and Canada for nonsmokers, cigarette smokers, and cigar and pipe smokers. The treatment factor here involves the three levels of smoking. It appears from the death rates in Part A of Table 1 that cigarette smoking is good for health, especially relative to cigar and pipe smoking, clearly a result contrary to current wisdom. A problem with the naive conclusion from Part A is exposed in Part B of Table 1, which gives the average ages of the subpopulations: age is correlated with both death rates and smoking behavior. Age in this example is a "confounding" covariate, and conclusions regarding the effects of smoking should be adjusted for differences in age distributions across subpopulations. A straightforward way of adjusting for age is to: (1) divide the population into age categories of approximately equal size (e.g., 2 categories = younger, older; or 3 categories = young, middle-age, old; or 4 categories, etc.); (2) compare death rates within an age category (e.g., within the younger population, compare death rates for the three treatment groups and similarly for the older population); and (3) average over the age-group-specific comparisons to obtain overall estimates of the age-adjusted death rates per 1000 for each of the three treatment groups. Part C of Table 1 shows the results for different numbers of categories of age, where the subclass age boundaries were defined to have equal numbers of nonsmokers in each subclass. These results, especially with 9-11 subclasses, align better than Part A with our current understanding of the effects of smoking. Incidentally, having approximately equal numbers of nonsmokers within each subclass is not necessary, but if the nonsmokers are considered the baseline group, it is a convenient and efficient choice because then the overall estimated effect is the simple unweighted average of the subclass specific results. That is, the mortality rates in all three groups are being "standardized" (Finch, 1988) to the age distribution of nonsmokers as defined by their subclass counts. Cochran (1968) calls this method "subclassification" and offers theoretical results showing that as long as the treatment groups overlap in their age distributions (i.e., as long as there are reasonable numbers of subjects from each treatment condition in each subclass), comparisons using 5 or 6 subclasses will typically remove 90% or more of the bias present in the raw comparisons in Part A. More than five subclasses were used in the final rows of Part C in Table 1 because the large sizes of the data sets made it possible to do so.

A particular statistical model such as a linear regression (or a logistic regression, or in other settings a hazard model), could have been used to adjust for age, but subclassification has two distinct advantages over such models, at least for offering initial trustworthy comparisons that are easy to communicate.

Tabel 1: Comparing Death Rates for Three Smoking Groups in each of Three Data Bases from Tables 1-3 in Cochran (1968)

	Canadian Study			UK Study			US Study		
	No Smoke	Cigarette	Cigar & Pipe	No Smoke	Cigarette	Cigar & Pipe	No Smoke	Cigarette	Cigar & Pipe
A	Death Rates per 1,000 Person Years								
	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
B	Average Age in Years								
	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
C	Adjusted Death Rates Using K Subclasses								
K=2	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
K=3	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
K=9-11	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

First, if the treatment groups do not adequately overlap on the confounding covariate age, the investigator will see it immediately and be warned. Thus, if members of one treatment group have ages outside the range of another group's ages, it will be obvious, because one or more age-specific subclasses will consist solely of members exposed to one treatment (or nearly so). In contrast, there is nothing in the standard output of any regression modeling software that will display this critical fact. The reason for this apparent omission is that such models predict an outcome (e.g., mortality) from regressors (e.g., age and treatment indicators), and standard regression diagnostics do not include the careful analysis of the joint distribution of the regressors (e.g., a comparison of the distributions of age across treatment groups). When the overlap on age distributions across treatment groups is too limited, the data base, no matter how large, cannot support causal conclusions about the differential effects of the treatments. For an extreme example, if the data base consists of 70 year-old smokers and 40 year-old nonsmokers, the comparison of 5-year survival rates among 70 year-old smokers and 40-year old nonsmokers provides essentially no information about the effect of smoking versus nonsmoking for either 70 year-olds or 40-year olds, or any other age group.

The second reason for preferring subclassification to models concerns more promising situations like that in Table 1, where the treatment groups overlap enough on the confounding covariate so that a comparison is possible. When estimating the treatment effect, subclassification does not rely on any particular functional form (e.g., linearity) for the relationship between the outcome (mortality) and covariate (age) within each treatment group, whereas models do rely on such assumptions. If the treatment groups have similar distributions of the covariate, common assumptions like linearity are usually harmless, but when the treatment groups have rather different covariate distributions, model-based methods of adjustment are dependent on the specific form of the model (e.g., linearity, log-linearity), and their answers are influenced by untrustworthy extrapolations. Simulations documenting the fragility of linear regression methods appear in Rubin (1973) for the case of one covariate.

If standard models can be so dangerous, why are they so commonly used for such adjustments when examining data bases for estimates of causal effects? One reason is the ease of automatic data analysis using existing, pervasive software on plentiful, speedy hardware. Nevertheless, although standard modeling software can automatically "handle" many regressor variables and produce results,

Mechanics Part Two: Subclassification / Stratification

- Propensity scores permit subclassification on multiple covariates simultaneously.
  - Permits the use of the whole sample of data (not just matched sets), without relying (as in regression adjustment) on a functional form
- Examples
  - Surgery vs. Medicine for Coronary Artery Disease
  - Security National Insurance
  - Weighting as a form of Stratification

Seminal paper: Rosenbaum and Rubin (1984)

Cochran's Subclassification Example: U.S. Male Death Rates per 1000 person-years

Smoking Group	Mean age in years	Unadjusted death rate	Adjusted for age (Two subclasses)
Non-smokers	57.0	13.5	13.5
Cigarettes only	53.2	13.5	16.4
Cigars, pipes	59.7	17.4	14.9

- Combine "low age" mortality rate in each smoking group with "high age" mortality rate in that group, weighting by population proportions of "low age" and "high age" U.S. males.

Cochran (1968, 1983), Rosenbaum and Rubin (1984)

Subclassification (Cochran) Smoking example

For each country (eg USA) subdivide non smokers (baseline group) age distrib. (confounder)

into 2 categories (median split)

or 5 categories (90% bias reduction) see lab4 bin(20th, 40th, 60th, 80th) percentiles

or >5

Equal numbers of non smokers in each bin
average over bins unchanged

Apply the binning to cigs and cigar

For cigs in USA younger than nonsmokers, so averaging over bins will overweight older cig smokers (higher mortality)

ave mortality over bin means moves 13.5 to 16.4 for 2 bins
(e.g. split USA cigs at med $\hat{=}$ 57 might yield 2/3, 1/3 division of subjects see HW8

use cut [Lab 4] or subset to form bins

Variable	Canadian Study			United Kingdom Study			United States Study		
	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers
Mortality rates per 1000 person-years, %	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
Average age, *y*	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
Adjusted mortality rates using subclasses, %									
2 subclasses	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3 subclasses	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9–11 subclasses	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

* Adapted from Tables 1–3 in Cochran (2).

understanding of the effects of smoking. Incidentally, having approximately equal numbers of nonsmokers within each subclass is not necessary, but if the nonsmokers are considered the baseline group, it is a convenient and efficient choice because then the overall estimated effect is the simple unweighted average of the subclass specific results. That is, the mortality rates in all three groups are being "standardized" (Finch, 1988) to the age distribution of nonsmokers as defined by their subclass counts.

Cochran (1968) calls this method "subclassification" and offers theoretical results showing that as long as the treatment groups overlap in their age distributions (i.e., as long as there are reasonable numbers of subjects from each treatment condition in each subclass), comparisons using 5 or 6 subclasses will typically remove 90% or more of the bias present in the raw comparisons in Part A. More than 5 subclasses were used in the final rows of Part C in Table 1 because the large sizes of the data sets made it possible to do so.

A particular statistical model such as a linear regression (or a logistic regression, or in other settings a hazard model), could have been used to adjust for age, but subclassification has two distinct advantages over such models, at least for offering initial trustworthy comparisons that are easy to communicate.

First, if the treatment groups do not adequately overlap on the confounding covariate age, the investigator will see it immediately and be warned. Thus, if members of one treatment group have ages outside the range of another group's ages, it will be obvious, because one or more age-specific subclasses will consist solely of members exposed to one treatment (or nearly so). In contrast, there is nothing in the standard output of any regression modelling software that will display this critical fact. The reason for this apparent omission is that such models predict an outcome (e.g., mortality) from regressors (e.g., age and treatment indicators), and standard regression diagnostics do not include the careful analysis of the joint distribution of the regressors (e.g., a comparison of the distributions of age across treatment groups). When the overlap on age distributions across treatment groups is too limited, the data base, no matter how large, cannot support causal conclusions about the differential effects of the treatments. For an extreme example, if the data base consists of 70 year-old smokers and 40 year-old nonsmokers, the comparison of 5-year survival rates among 70 year-old smokers and 40-year old nonsmokers provides essentially no information about the effect of smoking versus nonsmoking for either 70 year-olds or 40-year olds, or any other age group.

The second reason for preferring subclassification to models concerns more promising situations like that in Table 1, where the treatment groups overlap enough on the confounding covariate so that a comparison is possible. When estimating the treatment effect, subclassification does not rely on any particular functional form (e.g., linearity) for the relationship between the outcome (mortality) and covariate (age) within each treatment group, whereas models do rely on such assumptions. If the treatment groups have similar distributions of the covariate, common assumptions like linearity are usually harmless, but when the treatment groups have rather different covariate distributions, model-based methods of adjustment are dependent on the specific form of the model (e.g., linearity, log-linearity), and their answers are influenced by untrustworthy extrapolations. Simulations documenting the fragility of linear regression methods appear in Rubin (1973) for the case of one covariate.

If standard models can be so dangerous, why are they so commonly used for such adjustments when examining data bases for estimates of causal effects? One reason is the ease of automatic data analysis using existing, pervasive software on plentiful, speedy hardware. Nevertheless, although standard modelling software can automatically "handle" many regressor variables and produce results, these results can be remarkably misleading. In fact, when there are many confounding covariates, the issues of lack of adequate overlap and reliance on untrustworthy model-based extrapolations are even more serious than with only one confounding covariate, as documented by simulations in Rubin (1979, Table 2). One reason for the increased problem is that small differences on many covariates can accumulate into a substantial overall difference. For example, if one treatment group is a little older, has a little higher cholesterol, has a little more familial history of cancer, and so on, that group may be substantially less healthy. Another reason for the increased problem with many covariates rather than one covariate is that diagnosing nonlinear relationships between outcomes and many covariates is more complicated. Moreover, standard comparisons of means between the groups, like those in Table 1B, or even comparisons of histograms for each confounding covariate between the treatment groups, although adequate with one covariate, are inadequate with more than one. The groups may differ in a multivariate direction to an extent that cannot be discerned from separate analyses of each covariate. This multivariate direction is closely related to the statistical concept of the "best linear discriminant" and intuitively is the single combination of the covariates on which the treatment groups are farthest apart.

A second reason for the dominance of modelling over subclassification is the seeming difficulty of using subclassification when many confounding covariates, rather than one, need adjustment, which is the common case. Fortunately, subclassification techniques can be applied with many covariates with nearly the same reliability as with only one covariate. The key idea is to use "propensity score" techniques introduced by Rosenbaum and Rubin (1983a); these can be viewed as important extensions of discriminant matching techniques, which calculate the best linear discriminant between the treatment groups and match on it (Rubin, 1980). Since their introduction a decade and a half ago, propensity score methods have been used in a variety of applied problems in medical and other research disciplines (Aiken, Smith and Lake, 1994; Connors et alia, 1996; Cook and Goldman, 1988; Cook and Goldman, 1989; Drake and Fisher, 1995; Eastwood and Fisher, 1988; Fiebach et alia, 1990; Gu and Rosenbaum, 1993; Harrell et alia, 1990; Kane et alia, 1991; Lavori and Keller, 1988; Lavori, Keller and Endicott, 1988; Malloy et alia, 1990; Myers et alia, 1987; Reinisch, Sanders, Mortensen and Rubin, 1995; Rosenbaum and Rubin, 1984; Rosenbaum and Rubin, 1985a; Stone et alia, 1995; Willoughby et alia, 1990;). Nevertheless, propensity score methods have not been used nearly as frequently as they should have been relative to model-based methods.

STANFORD UNIVERSITY

obs studies checklist

1) Limit outcome comparison to regions (persons) where you have data [interp ~~extrap~~] (common support / subsamples)

2) Use X's for the purpose they are intended (correct, redress imbalance from self-selection etc)

3) Use X's in a robust manner (not ~~sensitive~~ sensitive (much) to what's included, measurement, funct form) i.e. fit rather than coeff.

4) Be able to see how well the X's function to restore balance (and remediate)

Don't use Y / reserve outcomes

for as if by experiment statement

recreate the hypothetical experiment

1.5
1.38

For HW8

From anderson et al

Example 6.5 Multivariate caliper matching: Consider a hypothetical study comparing two therapies effective in reducing blood pressure, where the investigators want to match on three variables: previously measured diastolic blood pressure, age, and sex. Such confounding variables can be divided into two types: categorical variables, such as sex, for which the investigators may insist on a perfect match (e = 0); and numerical variables, such as age and blood pressure, which require a specific value of the caliper tolerances. Let the blood pressure tolerance be specified as 5 mm Hg and the age tolerance as 5 years. Table 6.6 contains measurements of these three confounding variables. (The subjects are grouped by sex to make it easier to follow the example.)

Table 6.6 Hypothetical Measurements on Confounding Variables

Treatment Group				Comparison Reservoir			
Subject Number	Diastolic Blood Pressure (mm Hg)	Age	Sex	Subject Number	Diastolic Blood Pressure (mm Hg)	Age	Sex
1	94	39	F	1	80	35	F
2	108	56	F	2	120	37	F
3	100	50	F	3	85	50	F
4	92	42	F	4	90	41	F
5	65	45	M	5	90	47	F
6	90	37	M	6	90	56	F
				7	108	53	F
				8	94	46	F
				9	78	32	F
				10	105	50	F
				11	88	43	F
				12	100	42	M
				13	110	56	M
				14	100	46	M
				15	100	54	M
				16	110	48	M
				17	85	60	M
				18	90	35	M
				19	70	50	M
				20	90	49	M

curse of dimensionality, many matching variables

Costs of nuclear plants

A small comparative study from a classic text

`help.matchit` *HTML Help for Matchit Commands and Models*

Description

The `help.matchit` command launches html help for Matchit commands and supported methods. The full manual is available online at `http://gking.harvard.edu/matchit`.

Usage

```
help.matchit(object)
```

Arguments

`object`	a character string representing a Matchit command or model. `help.matchit("command")` will take you to an index of Matchit commands and `help.matchit("method")` will take you to a list of matching methods. The following inputs are currently available: `exact`, `nearest`, `subclass`, `full`, `optimal`.

Author(s)

Daniel Ho <⟨daniel.ho@yale.edu⟩>; Kosuke Imai <⟨kimai@princeton.edu⟩>; Gary King <⟨king@harvard.edu⟩>; Elizabeth Stuart<⟨stuart@stat.harvard.edu⟩>

See Also

The complete document is available online at `http://gking.harvard.edu/matchit`.

Lab 4 data for matching using MatchIt Is job training effective???

`lalonde` *Data from National Supported Work Demonstration and PSID, as analyzed by Dehejia and Wahba (1999).*

Description

This is a subsample of the data from the treated group in the National Supported Work Demonstration (NSW) and the comparison sample from the Current Population Survey (CPS). This data was previously analyzed extensively by Lalonde (1986) and Dehejia and Wahba (1999). The full dataset is available at `http://www.columbia.edu/~rd247/nswdata.html`.

Usage

```
data(lalonde)
```

From Lab 4

> dim(lalonde)□□[1] 614 10□□

> names(lalonde) "treat" "age" "educ" "black" "hispan" "married" "nodegree" "re74" "re75" "re78"

> attach(lalonde) > table(treat) treat□□ 0 1 □□429 185 □□

> lalonde[1:10,] treat age educ black hispan married nodegree re74 re75 re78

Format

614 actually

A data frame with 313 observations (185 treated, 429 control). There are 10 variables measured for each individual. "treat" is the treatment assignment (1=treated, 0=control). "age" is age in years. "educ" is education in number of years of schooling. "black" is an indicator for African-American (1=African-American, 0=not). "hispan" is an indicator for being of Hispanic origin (1=Hispanic, 0=not). "married" is an indicator for married (1=married, 0=not married). "nodegree" is an indicator for whether the individual has a high school degree (1=no degree, 0=degree). "re74" is income in 1974, in U.S. dollars. "re75" is income in 1975, in U.S. dollars. "re78" is income in 1978, in U.S. dollars.

Source

http://www.columbia.edu/~rd247/nswdata.html

References

Lalonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76: 604-620. \

Dehejia, R.H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs. *Journal of the American Statistical Association* 94: 1053-1062.

`match.data` *Output Matched Data Sets*

Description

`match.data` outputs matched data sets from `matchit()`.

Usage

```
match.data <- match.data(object, group="all", distance = "distance",
weights = "weights", subclass = "subclass")
```

Arguments

`object`	The output object from `matchit()`. This is a required input.
`group`	This argument specifies for which matched group the user wants to extract the data. Available options are `"all"` (all matched units), `"treat"` (matched units in the treatment group), and `"control"` (matched units in the control group). The default is `"all"`.
`distance`	This argument specifies the variable name used to store the distance measure. The default is `"distance"`.
`weights`	This argument specifies the variable name used to store the resulting weights from matching. The default is `"weights"`.
`subclass`	This argument specifies the variable name used to store the subclass indicator. The default is `"subclass"`.

Value

Returns a subset of the original data set sent to this-is-escaped-code{, with ju

The Lalonde Data

For all of our examples, we use data from the job training program analyzed in Lalonde (1986) and Dehejia & Wahba (1999). A subsample of the data consisting of the National Supported Work Demonstration (NSW) treated group and the comparison sample from the Population Survey of Income Dynamics (PSID) is included in MATCHIT, and the full dataset is available at `http://www.columbia.edu/~rd247/nswdata.html`.[5]

The variables in this dataset are in Table 1 below. One causal effect of interest is the impact that participation in the job training program, `treat==1`, had on real earnings in 1978, `re78`, for those that participated in the program, i.e., the average treatment effect on the treated (ATT):

(1)

where `re78(treat=1)` represents the potential outcome under the treatment of the job program, and `re78(treat=0)` under control. To be clear, note that the first term (inside the expectation) in Equation 1 is *observed*, whereas the second term is the *unobserved* counterfactual of real earnings if participants had not participated. The nature of causal inference is that one of the two terms in the difference will always be unobserved. The same expression of the ATT, in mathematical notation is:

(2)

Table 1: Description of Lalonde data

Name	Description
Outcome ( )	
`re78`	Real earnings (1978)
Treatment Indicator ( )	
`treat`	Treated in job training program from March 1975-June 1977 (1 if treated, 0 if not treated)
Pre-treatment Covariates ( )	
`age`	Age
`educ`	Years of education
`black`	Race black (1 if black, 0 otherwise)
`hispan`	Race hispanic (1 if Hispanic, 0 otherwise)
`married`	Marital status (1 if married, 0 otherwise)
`nodegree`	High school degree (1 if no degree, 0 otherwise)
`re74`	Real earnings (1974)
`re75`	Real earnings (1975)

Gary King 2005-03-09

Propensity Score Methods

- Rosenbaum and Rubin. "The Central Role of the Propensity Score in Observational Studies." Biometrika 1983.
- Observational study analogue of complete randomization
- The propensity score is the probability of treatment versus control as a function of observed covariates
  - Model the reasons for treatment versus control at the level of the decision makers
  - For example, logistic regression model to predict cigarette versus cigar/pipe smoking with age, education, income, etc. as predictors
- Then subclassify (or match) on the propensity score as if it were the only covariate, e.g., 5-10 subclasses
- If correctly done, this creates balance within each subclass on **ALL** covariates used to estimate the propensity score

Example: GAO Study of Breast Conservation versus Mastectomy

- Six large and expensive randomized clinical trials had been completed showing little difference for the type of women randomized in the trials and participating clinics
- Question: Same results in U.S. general practice?
- Observational data available
  - SEER Database: covariates, treatments, post-surgery outcomes
- Design phase
  - Hide outcomes
  - Think hard about decision rules and key covariates
  - Key covariates for decisions by doctors/women: Age, marital status, region of country, urbanization, race, size of tumor, etc., all available in SEER and considered sufficient
  - Balance covariates between treatment and control using subclasses

Estimated 5-year Survival Rates for Node-negative Patients in Six Randomized Clinical Trials

Study	Women		Estimated Survival Rate for Women		Estimated Causal Effect
	Breast Conservation (BC)	Mastectomy (Mas)	BC	Mas	BC – Mas
	n	n	%	%	%
US-NCI†	74	67	93.9	94.7	-0.8
Milanese†	257	263	93.5	93.0	0.5
French†	59	62	94.9	96.2	-1.3
Danish‡	289	288	87.4	85.9	1.5
EORTC‡	238	237	89.0	90.0	-1.0
US-NSABP‡	330	309	89.0	88.0	1.0

†Single-center trial; ‡ Multicenter trial

Reference: Rubin DB. Estimated Causal Effects from Large Datasets Using Propensity Scores. Annals of Internal Medicine 1997; 127, 8(II):757-763.

Estimated 5-year Survival Rates for Node-Negative Patients in the SEER Database within Each of Five Propensity Score Subclasses

Propensity Score Subclass	Women		Estimated Survival Rate for Women		Estimated Causal Effect
	Breast Conservation (BC)	Mastectomy (Mas)	BC	Mas	BC – Mas
	n	n	%	%	%
1	56	1008	85.6	86.7	-1.1
2	106	964	82.8	83.4	-0.6
3	193	866	85.2	88.8	-3.6
4	289	978	88.7	87.3	1.4
5	462	604	89.0	88.5	0.5
Averages Across Five Subclasses			86.3	86.9	-0.6

Reference: Rubin DB. Estimated Causal Effects from Large Datasets Using Propensity Scores. Annals of Internal Medicine 1997; 127, 8(II):757-763.

results on handout

Example - Propensity Subclassification

Several years ago the U.S. Government Accounting Office (GAO, 1994) summarized results from randomized experiments comparing mastectomy (removal of breast, but not the pectoral muscle, plus nodal dissection but no radiation) and breast-conservation therapy (lumpectomy, nodal dissection and radiation) for the treatment of breast cancer for node-negative patients. Table 2 is adopted from their Table 2, and the results there provide no evidence of any differential treatment effect, at least for the type of women who participated in these informed-consent clinical trials and received the kind of care dispensed at the centers participating in these trials. The question remained, however, how broadly these results could be generalized, i.e., to other node-negative women and other medical facilities. The GAO used the National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) observational data base to address this question. Restrictions (e.g., node-negative diagnosis, age 70 or younger, tumor 4 cm or smaller, etc., as detailed in GAO (1994) in its Tables 4 and I.3) were applied to correspond to criteria for the randomized experiments, and these reduced the data base to 1,106 women receiving breast-conservation therapy and 4,220 receiving mastectomy.
GAO used propensity score methods on the SEER database to compare the two treatments for breast cancer. First, approximately 30 potential confounding covariates and interactions were identified: year of diagnosis (1983-1985), age category (4 levels), tumor size, geographical registry (9 levels), race (4 levels), marital status (4 levels), and interactions of year and registry. A logistic regression was then used to predict treatment (mastectomy versus conservation therapy) from these confounding covariates based on the data from the 5,326 (1,106 + 4220) women. Each woman was then assigned an estimated propensity score -- her estimated probability, based on her covariate values, of receiving breast conservation therapy rather than mastectomy. The group of 5,326 was then divided into 5 approximately equal-size subclasses based on their individual propensity scores, just as if these propensity scores comprised the only covariate: 1,064 were in the most mastectomy-oriented subclass, 1,070 in the next subclass, 1,059 in the middle subclass, 1,067 in the next subclass, and 1,066 were in the most breast-conservation-oriented subclass.
Before examining any outcomes (i.e., any 5-year survival results) — and the "before" is critical, the subclasses were checked for balance on the covariates. Recall that propensity score theory claims that if the propensity scores are relatively constant within each subclass, then within each subclass, the distribution of all covariates should be approximately the same in both treatment groups. This balance was found to be satisfactory. If important within-subclass differences between treatment groups had been found on some covariates, then either the propensity score prediction model would need to be reformulated, or it would have been concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates. This process of cycling between checking for balance on the covariates and reformulating the propensity score model is described in Rosenbaum and Rubin (1984) in the context of a study investigating coronary bypass surgery. For example, if the variances of an important covariate were found to differ importantly between treatment and control groups, then the square of that covariate would have been included in the revised propensity score model. For another example, if the correlations between two important covariates differed between the groups, then the product of the covariates would have been added to the propensity score model.

Propensity scores were estimated by logistic regression, and they were used to create five subclasses of treatment/control women. The women were ranked by their estimated propensity scores, and the lowest 20% formed subclass 1, the next 20% formed subclass 2, etc. Within each subclass, balance was checked, not only on the covariates included in the propensity score, but also on all other important covariates in the database. For example, the average age of a treated women within each subclass should be approximately the same as the average age of a control women in that subclass, and the proportion of each that are married should also be as similar as if the treatment and control women in that subclass had been randomly divided (obviously, not with equal probability across the subclasses). When less balance was found on a key covariate within a subclass than would have occurred in a randomized experiment, terms were added to the propensity score model and balance was reassessed. Unfortunately, those tables and the processes never survived into the final report, but such balance was achieved—not perfectly, but close enough to believe in the hypothetical underlying randomized block experiment that led to the observed data.

The results of the subclassification on the propensity score are summarized in Table 2. In general, this observational study's results are consistent with those from the randomized trials. There is essentially no evidence for any advantage to the radical operation, except possibly in those propensity score subclasses where the women and doctors were more likely to select mastectomy (subclasses 1, 2, 3), but the data are certainly not definitive. Similarly, for the women and doctors relatively more likely to select breast conserving operations, there is some slight evidence of a survival benefit to that choice. If we believed that the treatment effect should be the same for all women in the study, these changing results across propensity subclasses could be viewed as evidence of a confounded and nonignorable treat-

TABLE 2

Estimated 5-year survival rates for node-negative patients in SEER data base within each of five propensity score subclasses: from tables in U.S. GAO Report [*General Accounting Office* (1994)]

Propensity score subclass	**Treatment condition**	***n***	**Estimate**
1	Brest conservation	56	85.6%
	Mastectomy	1008	86.7%
2	Brest conservation	106	82.8%
	Mastectomy	964	83.4%
3	Brest conservation	193	85.2%
	Mastectomy	866	88.8%
4	Brest conservation	289	88.7%
	Mastectomy	978	87.3%
5	Brest conservation	462	89.0%
	Mastectomy	604	88.5%

AppA, RR 1984

Stat 209

Week 8 Propensity Scores

Let $z = 1, 0$ T/C $\underset{\sim}{X}$ vector of covariates

propensity score $e(\underset{\sim}{X}) = Pr(z=1 | \underset{\sim}{X})$ scalar $\hat{e}(\underset{\sim}{X})$

cond'l prob unit w vector $\underset{\sim}{X}$ observed cov. assigned to T ($z=1$)

Thm Balancing score $b(\underset{\sim}{X})$ s.t. conditional distrib of $\underset{\sim}{X}$ given $b(\underset{\sim}{X})$ same of treated and control units

$\underset{\sim}{X} \perp\!\!\!\perp z | b(\underset{\sim}{X})$. Coarsest (low dimen) balancing score is propensity score. $Pr(X, z|e) = Pr\{\underset{\sim}{X}|e\} Pr(z|e)$

Thm (result) Approx 90% reduction in bias for subclassifying at quintiles of population propensity score. $B_I = E(f(\underset{\sim}{x})|z=1) - E(f(\underset{\sim}{x})|z=0)$, B_S after strat, treatm percent reduction in bias $100(1 - B_S/B_I) \approx 90\%$

Ros Rubin
1983 Biometrika
1984 JASA

(i) The propensity score is a balancing score.
(ii) Any score that is 'finer' than the propensity score is a balancing score; moreover, x is the finest balancing score and the propensity score is the coarsest.
(iii) If treatment assignment is strongly ignorable given x, then it is strongly ignorable given any balancing score
(iv) At any value of a balancing score, the difference between the treatment and control means is an unbiased estimate of the average treatment effect at that value of the balancing score if treatment assignment is strongly ignorable. Consequently, with strongly ignorable treatment assignment, pair matching on a balancing score, subclassification on a balancing score and covariance adjustment on a balancing score can all produce unbiased estimates of treatment effects.
(v) Using sample estimates of balancing scores can produce sample balance on x.

Applications: Rubin Breast Cancer, Love (RR '84) CAD, Love Aspirin, Hansen SAT coaching, Substance Abuse (UNC), Rosenbaum, Danish downers

Rubin Ann Int Medicine

BC -> T

Lalonde data

Lab 4 stratification

Table 3: Estimated 5-year Survival Rates for Node-Negative Patients in SEER from Tables 5 and 7 in U.S. GAO Report (1994). AIM pub

Propensity Score

Subclass	Treatment	n	Estimate	n*	Estimate*
1	Breast Conservation	56	85.6%	54	88.8%
	Mastectomy	1,008	86.7%	966	90.5%
2	Breast Conservation	106	82.8%	102	86.0%
	Mastectomy	964	82.8%	917	87.7%
3	Breast Conservation	193	85.2%	184	89.4%
	Mastectomy	866	88.8%	841	91.4%
4	Breast Conservation	289	88.7%	279	92.0%
	Mastectomy	978	87.3%	742	91.5%
5	Breast Conservation	462	89.0%	453	90.7%
	Mastectomy	604	88.5%	589	90.7%

* omitting patients whose deaths were unrelated to cancer.

```
> table(propbin, treat)
                 treat
propbin            0   1
  (0,0.0401]     122   1
  (0.0401,0.0872] 116   7
  (0.0872,0.27]  101  21
  (0.27,0.671]    53  71
  (0.671,1]       37  85
> tapply(re78, list(propbin, treat),mean)
                    0    1
(0,0.0401]      10467    0
(0.0401,0.0872]  5797 7919
(0.0872,0.27]    6043 9211
(0.27,0.671]     4977 5819
(0.671,1]        4666 6030
```

counts

means re78

Matching in Statistics: Cochran's School in the 1980s

- Propensity score
  - Close matches on multivariate **x** not needed if you can match closely on scalar $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1983, 1984).
  - Good to combine matching on **x** with matching on $\phi(\mathbf{x})$, privileging closeness on $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1985).
- Computerized matching $\rightarrow$ optimal matching (Rosenbaum, 1989)

General Procedure

Run Logistic Regression:

- Dependent variable: Y=1, if participate; Y = 0, otherwise.
- Choose appropriate conditioning (instrumental) variables.
- Obtain propensity score: predicted probability (p) or log[(1-p)/p].

Either

- 1-to-1 or 1-to-n match and then stratification (subclassification)
- Kernel or local linear weight match and then estimate Difference-in-differences (Heckman)

Or

1-to-1 or 1-to-n Match

- Nearest neighbor matching
- Caliper matching
- Mahalanobis
- Mahalanobis with propensity score added

Multivariate analysis based on new sample

LAB 4 excerpt

```
# now do the logistic regression that computes propensity scores (matching packages will do this for
> glm.lalonde = glm( treat ~ age + educ + black + hispan + married + nodegree + re74 + re75,
+ data = lalonde, family = binomial)
> propen = fitted(glm.lalonde)  # now we have the propensity scores, Lab script calls these propScore
> tapply(propen, treat, quantile) # look at overlap via 5-number summary (or side-by-side boxplots)
                                  not real good overlap, as noted in class handout
$`0`
     0%     25%     50%     75%    100%
0.00908 0.03888 0.07585 0.19514 0.78917

$`1`
     0%     25%     50%     75%    100%
0.02495 0.52646 0.65368 0.72660 0.85315

> # the common use of the propensity scores (backed by theory, class handout 2/26))
> # is to stratify by quintiles

> # the simple-minded way I do it is to use "cut", Lab script is fancier programming
> ?cut  # this is a simple function to create bins
> k = 1:4
> quantile(propen, k/5)
    20%     40%     60%     80%
0.04015 0.08721 0.26978 0.67085
>  propbin = cut(propen, c(0, .04015,.08721,.26978,.67085,1))


> table(propbin, treat) # either way you display it, we do not have good overlap in the bottom
                        two quintiles, lower estimated probability for being in treatment
                        for treatment cases
                 treat
propbin             0   1
  (0,0.0401]      122   1
  (0.0401,0.0872] 116   7
  (0.0872,0.27]   101  21
  (0.27,0.671]     53  71
  (0.671,1]        37  85


> tapply(re78, list(propbin, treat),mean)  # here are the mean diffs in re78 the outcome
                                           stratified by propensity quintile
# direction of mean diffs favors treatment, job training
                     0    1
(0,0.0401]       10467    0
(0.0401,0.0872]  5797 7919
(0.0872,0.27]    6043 9211
(0.27,0.671]     4977 5819
(0.671,1]        4666 6030


> t.test(re78[propbin == bins[5]] ~ treat[propbin == bins[5]]) # t-test for quintile 5

etc
```

Propensity Score Methods

- Rosenbaum and Rubin. "The Central Role of the Propensity Score in Observational Studies." Biometrika 1983.
- Observational study analogue of complete randomization
- The propensity score is the probability of treatment versus control as a function of observed covariates
  - Model the reasons for treatment versus control at the level of the decision makers
  - For example, logistic regression model to predict cigarette versus cigar/pipe smoking with age, education, income, etc. as predictors
- Then subclassify (or match) on the propensity score as if it were the only covariate, e.g., 5-10 subclasses
- If correctly done, this creates balance within each subclass on **ALL** covariates used to estimate the propensity score

Example: GAO Study of Breast Conservation versus Mastectomy

- Six large and expensive randomized clinical trials had been completed showing little difference for the type of women randomized in the trials and participating clinics
- Question: Same results in U.S. general practice?
- Observational data available
  - SEER Database: covariates, treatments, post-surgery outcomes
- Design phase
  - Hide outcomes
  - Think hard about decision rules and key covariates
  - Key covariates for decisions by doctors/women: Age, marital status, region of country, urbanization, race, size of tumor, etc., all available in SEER and considered sufficient
  - Balance covariates between treatment and control using subclasses

results on handout

Example - Propensity Subclassification

Several years ago the U.S. Government Accounting Office (GAO, 1994) summarized results from randomized experiments comparing mastectomy (removal of breast, but not the pectoral muscle, plus nodal dissection but no radiation) and breast-conservation therapy (lumpectomy, nodal dissection and radiation) for the treatment of breast cancer for node-negative patients. Table 2 is adopted from their Table 2, and the results there provide no evidence of any differential treatment effect, at least for the type of women who participated in these informed-consent clinical trials and received the kind of care dispensed at the centers participating in these trials. The question remained, however, how broadly these results could be generalized, i.e., to other node-negative women and other medical facilities. The GAO used the National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) observational data base to address this question. Restrictions (e.g., node-negative diagnosis, age 70 or younger, tumor 4 cm or smaller, etc., as detailed in GAO (1994) in its Tables 4 and I.3) were applied to correspond to criteria for the randomized experiments, and these reduced the data base to 1,106 women receiving breast-conservation therapy and 4,220 receiving mastectomy.
GAO used propensity score methods on the SEER database to compare the two treatments for breast cancer. First, approximately 30 potential confounding covariates and interactions were identified: year of diagnosis (1983-1985), age category (4 levels), tumor size, geographical registry (9 levels), race (4 levels), marital status (4 levels), and interactions of year and registry. A logistic regression was then used to predict treatment (mastectomy versus conservation therapy) from these confounding covariates based on the data from the 5,326 (1,106 + 4220) women. Each woman was then assigned an estimated propensity score -- her estimated probability, based on her covariate values, of receiving breast conservation therapy rather than mastectomy. The group of 5,326 was then divided into 5 approximately equal-size subclasses based on their individual propensity scores, just as if these propensity scores comprised the only covariate: 1,064 were in the most mastectomy-oriented subclass, 1,070 in the next subclass, 1,059 in the middle subclass, 1,067 in the next subclass, and 1,066 were in the most breast-conservation-oriented subclass.
Before examining any outcomes (i.e., any 5-year survival results) — and the "before" is critical, the subclasses were checked for balance on the covariates. Recall that propensity score theory claims that if the propensity scores are relatively constant within each subclass, then within each subclass, the distribution of all covariates should be approximately the same in both treatment groups. This balance was found to be satisfactory. If important within-subclass differences between treatment groups had been found on some covariates, then either the propensity score prediction model would need to be reformulated, or it would have been concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates. This process of cycling between checking for balance on the covariates and reformulating the propensity score model is described in Rosenbaum and Rubin (1984) in the context of a study investigating coronary bypass surgery. For example, if the variances of an important covariate were found to differ importantly between treatment and control groups, then the square of that covariate would have been included in the revised propensity score model. For another example, if the correlations between two important covariates differed between the groups, then the product of the covariates would have been added to the propensity score model.

Estimated 5-year Survival Rates for Node-negative Patients in Six Randomized Clinical Trials

Study	Women		Estimated Survival Rate for Women		Estimated Causal Effect
	Breast Conservation (BC)	Mastectomy (Mas)	BC	Mas	BC – Mas
	n	n	%	%	%
US-NCI†	74	67	93.9	94.7	-0.8
Milanese†	257	263	93.5	93.0	0.5
French†	59	62	94.9	96.2	-1.3
Danish‡	289	288	87.4	85.9	1.5
EORTC‡	238	237	89.0	90.0	-1.0
US-NSABP‡	330	309	89.0	88.0	1.0

†Single-center trial; ‡ Multicenter trial

Reference: Rubin DB. Estimated Causal Effects from Large Datasets Using Propensity Scores. Annals of Internal Medicine 1997; 127, 8(II):757-763.

Estimated 5-year Survival Rates for Node-Negative Patients in the SEER Database within Each of Five Propensity Score Subclasses

Propensity Score Subclass	Women		Estimated Survival Rate for Women		Estimated Causal Effect
	Breast Conservation (BC)	Mastectomy (Mas)	BC	Mas	BC – Mas
	n	n	%	%	%
1	56	1008	85.6	86.7	-1.1
2	106	964	82.8	83.4	-0.6
3	193	866	85.2	88.8	-3.6
4	289	978	88.7	87.3	1.4
5	462	604	89.0	88.5	0.5
Averages Across Five Subclasses			86.3	86.9	-0.6

Reference: Rubin DB. Estimated Causal Effects from Large Datasets Using Propensity Scores. Annals of Internal Medicine 1997; 127, 8(II):757-763.

Propensity scores were estimated by logistic regression, and they were used to create five subclasses of treatment/control women. The women were ranked by their estimated propensity scores, and the lowest 20% formed subclass 1, the next 20% formed subclass 2, etc. Within each subclass, balance was checked, not only on the covariates included in the propensity score, but also on all other important covariates in the database. For example, the average age of a treated women within each subclass should be approximately the same as the average age of a control women in that subclass, and the proportion of each that are married should also be as similar as if the treatment and control women in that subclass had been randomly divided (obviously, not with equal probability across the subclasses). When less balance was found on a key covariate within a subclass than would have occurred in a randomized experiment, terms were added to the propensity score model and balance was reassessed. Unfortunately, those tables and the processes never survived into the final report, but such balance was achieved—not perfectly, but close enough to believe in the hypothetical underlying randomized block experiment that led to the observed data.

The results of the subclassification on the propensity score are summarized in Table 2. In general, this observational study's results are consistent with those from the randomized trials. There is essentially no evidence for any advantage to the radical operation, except possibly in those propensity score subclasses where the women and doctors were more likely to select mastectomy (subclasses 1, 2, 3), but the data are certainly not definitive. Similarly, for the women and doctors relatively more likely to select breast conserving operations, there is some slight evidence of a survival benefit to that choice. If we believed that the treatment effect should be the same for all women in the study, these changing results across propensity subclasses could be viewed as evidence of a confounded and nonignorable treat-

TABLE 2

Estimated 5-year survival rates for node-negative patients in SEER data base within each of five propensity score subclasses: from tables in U.S. GAO Report [*General Accounting Office* (1994)]

Propensity score subclass	**Treatment condition**	***n***	**Estimate**
1	Brest conservation	56	85.6%
	Mastectomy	1008	86.7%
2	Brest conservation	106	82.8%
	Mastectomy	964	83.4%
3	Brest conservation	193	85.2%
	Mastectomy	866	88.8%
4	Brest conservation	289	88.7%
	Mastectomy	978	87.3%
5	Brest conservation	462	89.0%
	Mastectomy	604	88.5%

Multivariate Matching with the Propensity Score

- Match subjects so that they balance on multiple covariates using one scalar score.
- Goal: Emulate a RCT in matching, then use standard analyses to compare matched sets.
- Design: Treated subjects matched to people who didn't receive treatment but who had similar propensity to receive treatment (match the treated to untreated "clones").

Aspirin Use and Mortality

- 6174 consecutive adults at CCF 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 et al. (2001) http://www.ncbi.nlm.nih.gov/pubmed/11559263

JAMA. 2001 Sep 12;286(10):1187-94. http://jama.jamanetwork.com/article.aspx?articleid=194177
Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. Gum PA1, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS.

Propensity Score Model for Aspirin Use

- Logistic Regression predicting aspirin use
- 31 covariates included in the model:
  - Demographics, Clinical history, Medication use
  - Cardiovascular assessment and Exercise capacity
- Estimated propensity scores for aspirin use range from .03 to .98
  - ROC Area shows good discrimination (C = .83)
- But does the propensity score model work?
- Are the covariates balanced?

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.

Matching with Propensity Scores

- For each patient, we have a propensity score.
- Randomly select an Aspirin user.
- Match to the non-user with closest propensity score (within some limit or "calipers")
- Eliminate both patients from pool, and repeat until you can't find an acceptable match.
  - Could match a non-user with Propensity Score inside "calipers" who matches exactly on characteristic X, or…
  - Match non-user with Propensity score inside "calipers" and smallest "distance" on some pre-specified covariates.

OR do a fullmatch or optimal match maybe wit restrictions. What would Ben do?

Matching on Gender within PS Calipers

- Shuffle "treatment" patients, and select one.
- Find all "non-treated" with PS inside calipers (here we'll set calipers at treated PS ± .03).
- Match patient within calipers of same gender.
- Repeat until no more matches are possible.

.80
.79
.78
.77
.76
.75
.74
.73
.72

Patient	Exposure	PS	Gender
A	**Treated**	**.76**	**Male**
B	Not Treated	.77	Female
C	Not Treated	.74	Male
D	Not Treated	.80	Male

How Were The Aspirin Subjects Matched?

- Tried to match each aspirin user to a unique non-user with a PS identical to 5 digits.
- If not possible, proceeded to a 4-digit match, then 3-digit, 2-digit, and finally a 1-digit match (i.e., propensity scores within .099).
- Result: matches for 1351 (58%) of the 2310 aspirin patients to 1351 unique non-users.

SAS macro: http://www2.sas.com/proceedings/sugi26/p214-26.pdf

Propensity Matcher Results

ID	Treated?	Propensity	Linear Propensity	Match?	Partner ID
1	1	0.2	-1.386	No	-999
2	1	0.3	-0.847	Yes	8
3	1	0.4	-0.405	Yes	10
4	1	0.6	0.405	No	-999
5	1	0.7	0.847	No	-999

logit

SE (Linear Propensity):	0.1829
x % Selected:	0.6
x % of SE:	0.1097

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]

Using Standardized Differences to Measure Covariate Balance

- Standardized Differences are appropriate summaries of Covariate Balance for both Continuous and Categorical Variables

$$d = \frac{100\left(\bar{x}_{Treatment} - \bar{x}_{Control}\right)}{\sqrt{\frac{s^2_{Treatment} + s^2_{Control}}{2}}}$$ for continuous variables

$$d = \frac{100\left(p_{Treatment} - p_{Control}\right)}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}}$$ for binary variables

|Standardized Differences| > 10% Indicate Serious Imbalance

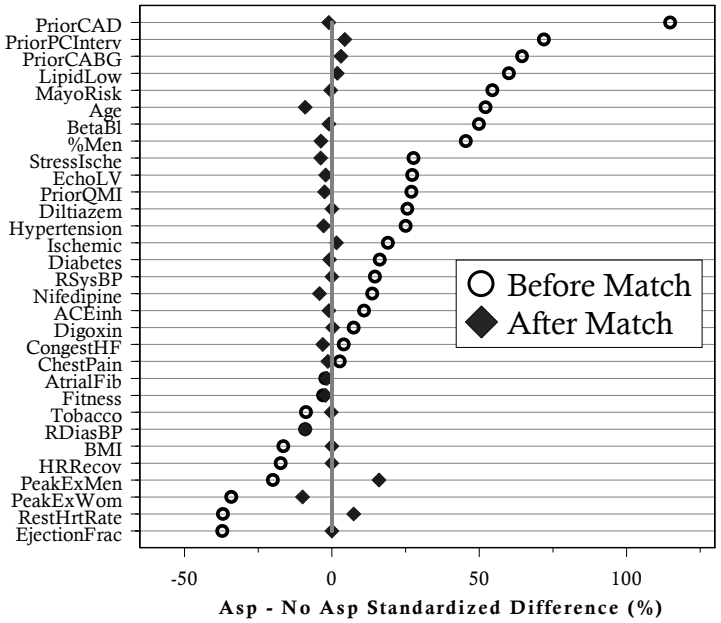
Before Match:

- 811/2310 (35.1%) Aspirin users used β-blockers
- 550/3864 (14.2%) non-Aspirin users used β-blockers
- Standardized Difference is 49.9%
- P value for difference is < .001

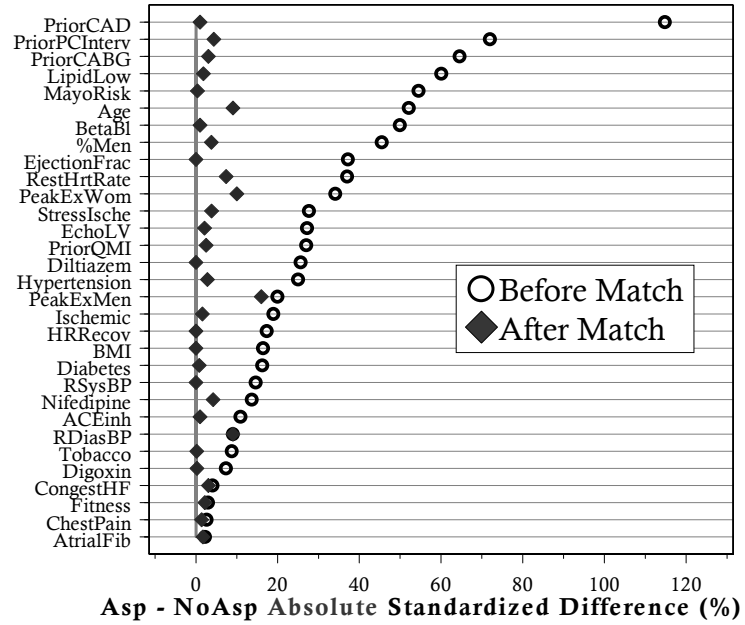
After Match:

- 352/1351 (26.1%) Aspirin users used β-blockers
- 358/1351 (26.5%) non-Aspirin users used β-blockers
- Standardized Difference is –1.0%
- P value for difference is .79

Covariate Balance for Aspirin Study



Absolute Standardized Differences



Matching with Propensity Scores

- 1351 aspirin subjects matched well to non-aspirin subjects – big improvement in covariate balance. Matched group looks like an RCT…
- Matching still incomplete, but results on PS matched group mirrored the results for the covariate-adjusted group as a whole…
- Resulting matched pairs analyzed using standard statistical methods, e.g. Kaplan-Meier, Cox proportional hazards models.

Estimating The Hazard Ratios

Approach	n	Hazard Ratio	95% CI
Full sample, no adjustment	6174	**1.08**	(.85, 1.39)
Full sample with no PS, adjusted for all covariates	6174	**0.67**	(.51, .87)
PS-Matched sample	2702	**0.53**	(.38, .74)
PS-Matched, adjusted for PS and all covariates	2702	**0.56**	(.40, .78)

- During follow-up 153 (6%) of the 2702 propensity score-matched patients died.
- Aspirin use was associated with a lower risk of death in matched group (4% vs. 8%, p = .002).

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease

A Propensity Analysis

Patricia A. Gum, MD
Maran Thamilarasan, MD
Junko Watanabe, MD
Eugene H. Blackstone, MD
Michael S. Lauer, MD

Context Although aspirin has been shown to reduce cardiovascular morbidity and short-term mortality following acute myocardial infarction, the association between its use and long-term all-cause mortality has not been well defined.

Objectives To determine whether aspirin is associated with a mortality benefit in stable patients with known or suspected coronary disease and to identify patient characteristics that predict the maximum absolute mortality benefit from aspirin.

Design and Setting Prospective, nonrandomized, observational cohort study conducted between 1990 and 1998 at an academic medical institution, with a median follow-up of 3.1 years.

Patients Of 6174 consecutive adults undergoing stress echocardiography for evaluation of known or suspected coronary disease, 2310 (37%) were taking aspirin. Patients with significant valvular disease or documented contraindication to aspirin use, including peptic ulcer disease, renal insufficiency, and use of nonsteroidal anti-inflammatory drugs, were excluded.

Main Outcome Measure All-cause mortality according to aspirin use.

Results During 3.1 years of follow-up, 276 patients (4.5%) died. In a simple univariable analysis, there was no association between aspirin use and mortality (4.5% vs 4.5%). However, after adjustment for age, sex, standard cardiovascular risk factors, use of other medications, coronary disease history, ejection fraction, exercise capacity, heart rate recovery, and echocardiographic ischemia, aspirin use was associated with reduced mortality (hazard ratio [HR], 0.67; 95% confidence interval [CI], 0.51-0.87; $P = .002$). In further analysis using matching by propensity score, 1351 patients who were taking aspirin were at lower risk for death than 1351 patients not using aspirin (4% vs 8%, respectively; HR, 0.53; 95% CI, 0.38-0.74; $P = .002$). After adjusting for the propensity for using aspirin, as well as other possible confounders and interactions, aspirin use remained associated with a lower risk for death (adjusted HR, 0.56; 95% CI, 0.40-0.78; $P<.001$). The patient characteristics associated with the most aspirin-related reductions in mortality were older age, known coronary artery disease, and impaired exercise capacity.

Conclusion Aspirin use among patients undergoing stress echocardiography was independently associated with reduced long-term all-cause mortality, particularly among older patients, those with known coronary artery disease, and those with impaired exercise capacity.

JAMA. 2001;286:1187-1194 www.jama.com

ASPIRIN HAS BEEN SHOWN TO be associated with decreased cardiovascular morbidity in multiple clinical trials[1,2] but the association between aspirin use and all-cause mortality has been less well defined except in the setting of acute myocardial infarction.[3] Although a few observational analyses have suggested a longer-term survival benefit,[4-6] it is not clear whether this benefit persists after accounting for treatment selection biases as well as established predictors of survival in patients with known or suspected coronary artery disease, in particular impaired exercise capacity, left ventricular dysfunction, and myocardial ischemia.

In this study we sought, based on an a priori hypothesis, to determine if aspirin use was associated with a reduction in all-cause mortality among stable patients referred for stress echocardiography. Because the validity of observational studies of treatment effects may be limited by selection biases and confounding factors, we performed a propensity analysis.[7]

For editorial comment see p 1228.

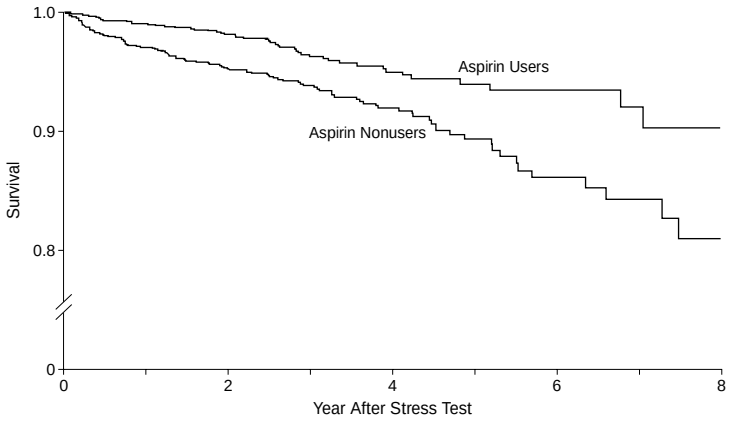
METHODS

Patients

The study sample was derived from 9954 consecutive adult patients undergoing stress echocardiography at the Cleveland Clinic Foundation between 1990

Author Affiliations: Departments of Cardiology (Drs Gum, Thamilarasan, Watanabe, and Lauer), Thoracic and Cardiovascular Surgery (Dr Blackstone), and Biostatistics and Epidemiology (Dr Blackstone), Cleveland Clinic Foundation, Cleveland, Ohio.
Corresponding Author and Reprints: Michael S. Lauer, MD, Department of Cardiology, Desk F25, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195 (e-mail: lauerm@ccf.org).

Figure 1. Kaplan-Meier Curve Relating Aspirin Use to Time to Death Among Propensity-Matched Patients



No. at Risk									
Aspirin Use	1351	1338	1328	1314	1308	1305	1304	1303	1300
No Aspirin Use	1351	1312	1292	1280	1269	1260	1254	1252	1249

Table 4. Cox Proportional Hazards Analyses of Aspirin Use and Mortality Among Propensity-Matched Patients (n = 2702)*

Model	Hazard Ratio (95% CI)	*P* Value
Unadjusted	0.53 (0.38-0.74)	.002
Adjusted for propensity	0.53 (0.38-0.74)	<.001
Adjusted for propensity and selected variables†	0.59 (0.42-0.83)	.002
Adjusted for propensity and all covariates‡	0.56 (0.40-0.78)	<.001

*CI indicates confidence interval.
†Selected variables included prior coronary artery disease, prior coronary artery bypass grafting, prior percutaneous intervention, and ejection fraction ≤40%.
‡For a list of covariates, see Table 2 footnote (†).

history,[1,25-27] patients with chronic stable angina,[28,29] patients presenting with AMI,[3,6] and patients with unstable angina.[30-33] Randomized trial evidence demonstrates that aspirin reduces all-cause mortality among patients with AMI.[3] It is less clear if aspirin use reduces long-term all-cause mortality in stable patient populations. Two recent observational analyses of patients enrolled in the Bezafibrate Infarction Prevention Trial demonstrated reduced mortality rates among patients taking aspirin, irrespective of the presence or absence of diabetes or therapy with ACE inhibitors.[4,5] Furthermore, the Collaborative Group of the Primary Prevention Project recently demonstrated in a randomized trial a similar, although not statistically significant, reduction in relative risk for all-cause mortality (0.81; 95% CI, 0.58-1.13).[34] These findings are similar to ours but did not reach statistical significance, most likely due to a small number of events.

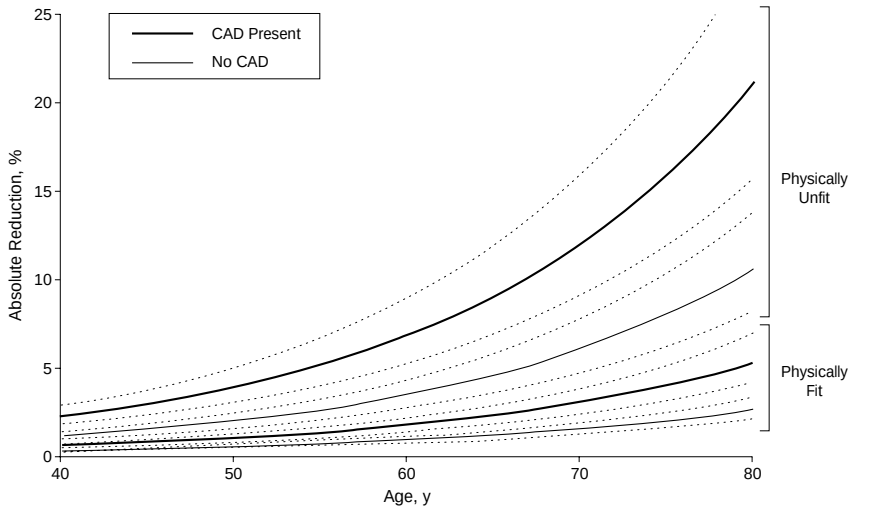
The current study extends these previous findings in several important respects. First, we demonstrated that aspirin use is associated with a reduction in long-term all-cause mortality, which is a clinically relevant, objective, and wholly unbiased end point.[16] Second, because we focused on patients referred for stress echocardiography we were able to account for several critical predictors of mortality, including left ventricular systolic function, stress-induced myocardial ischemia, and impaired exercise capacity. Third, unlike prior observational studies of aspirin use and outcome,[4,5,26] we used propensity analysis, which has been argued to be a powerful means of accounting for baseline confounding and selection biases.[7]

Furthermore, we observed this mortality reduction in a large cohort of consecutive patients seen within a clinical practice, as opposed to a clinical trial. It has been argued that patients enrolled in clinical trials may not be representative of patients seen in practice.[35] The patients included in our study population may represent a more representative sample of "real world" patients referred for evaluation of known or suspected cardiovascular disease than those included in many of the randomized controlled trials that have previously evaluated aspirin use for mortality reduction. Among the patients included in the Physicians' Health Study, 84% had no history of cardiovascular disease.[1] Additionally, those patients and those evaluated in other primary prevention trials had low rates of cardiovascular risk factors.[27,34] The studies evaluating aspirin use by patients with unstable angina also enrolled comparatively few patients with multiple cardiac risk factors or positive histories of previous coronary intervention.[31,33,36] Thus, the lower-risk population enrolled in the previous randomized trials may have contributed to their finding no mortality benefit. Furthermore, in a follow-up report of the Physicians' Health Study evaluating posttrial self-selected aspirin use and subsequent mortality, self-selected aspirin use was associated with multiple cardiovascular risk factors and a decrease in all-cause mortality.[37]

The mechanisms by which aspirin may reduce mortality include its platelet-blocking effects, its anti-inflammatory properties, or other as-yet unknown actions. Aspirin has been shown to be a powerful antiplatelet agent that acts by blocking the production of thromboxane A_2,[38] which may then reduce the risk of fatal cardiovascular events.[39,40] Recently, increasing interest has focused on inflammation, as assessed by C-reactive protein levels and cardiovascular risk.[41,42] Aspirin has been shown to reduce C-reactive protein levels.[41] In the randomized Physicians' Health Study the reduction in cardiovascular risk associated with aspirin was most pronounced among men with elevated baseline C-reactive protein levels.[43]

The major limitation of this study is that aspirin use was not based on a randomized assignment. Although the use of observational studies for assessment of treatment effects is controversial,[44] recent work has suggested that observational studies, when properly done, are not likely to produce misleading or bi-

Figure 2. Predicted Absolute Reduction in 5-Year Mortality by Age, Exercise Capacity, and History of CAD



Estimates are based on wholly parametric multivariable patient-specific survival equations. For each patient, equations were solved twice, once assuming aspirin use and once assuming nonuse. Dashed lines represent 95% confidence intervals. Methods used to derive these curves are explained in the "Methods" section and elsewhere.[23] CAD indicates coronary artery disease. Physically unfit is defined as fair or poor functional capacity for age and sex.[13]

ased results.[45-47] Furthermore, we used propensity analysis to enable an even more rigorous adjustment for selection bias and confounding than would be possible with standard multivariable analysis.[7] Nonetheless, it must be acknowledged that observational studies can only partially control for factors actually measured and can adjust for these factors only as well as the instrument used to measure them is capable. In contrast, randomization allocates both known and unknown confounding variables and avoids the introduction of bias from either the participants or their physicians. Other limitations of our study included lack of information about aspirin dose, aspirin allergy, or duration of treatment, as well as lack of data regarding medication adjustments made after stress testing.

Despite these limitations, the association between aspirin use and reduced mortality meets currently accepted criteria for likely causality.[48] The association was strong, with a greater than 30% reduction in risk of death. A temporal pattern is evident in Kaplan-Meier analyses. Biological plausibility is present, considering the known importance of increased platelet activity associated with coronary artery disease, aging,[49] and impaired physical fitness.[24] Our results are consistent with other observational non–propensity-adjusted analyses[5] and with a recent randomized study,[34] and the association appears to be largely unaffected by possible bias and confounding, whether assessed by standard multivariable analyses or more rigorous propensity analyses. Thus, our findings provide additional support for recommending the routine use of aspirin in patients with, or at risk for, cardiovascular disease—not only for preventing morbid events but also for reducing all-cause mortality.

Author Contributions: *Study concept and design:* Gum, Thamilarasan, Lauer.
Acquisition of data: Thamilarasan, Watanabe, Lauer.
Analysis and interpretation of data: Gum, Blackstone, Lauer.
Drafting of the manuscript: Gum, Lauer.
Critical revision of the manuscript for important intellectual content: Gum, Thamilarasan, Watanabe, Blackstone, Lauer.
Obtained funding: Lauer; Blackstone.
Statistical expertise: Blackstone, Lauer.
Administrative, technical, or material support: Thamilarasan, Watanabe, Lauer.
Study supervision: Lauer.

Funding/Support: Drs Lauer and Blackstone receive support from the American Heart Association (grant 0040244N) and from the National Heart, Lung, and Blood Institute (grant HL 66004-01). None of the investigators own stock, equity, or receive any form of remuneration from any pharmaceutical or medical device company.
Acknowledgment: We are grateful to Lori Parsons of Ovation Research Group, Seattle, Wash, for providing us with the SAS macro for propensity matching and for her advice regarding its use.

REFERENCES

1. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med*. 1989;321:129-135.
2. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy, I: prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *BMJ*. 1994;308:81-106.
3. Second International Study of Infarct Survival (ISIS-2) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*. 1988;2:349-360.
4. Leor J, Reicher-Reiss H, Goldbourt U, et al. Aspirin and mortality in patients treated with angiotensin-converting enzyme inhibitors: a cohort study of 11,575 patients with coronary artery disease. *J Am Coll Cardiol*. 1999;33:1920-1925.
5. Harpaz D, Gottlieb S, Graff E, Boyko V, Kishon Y, Behar S, for the Israeli Bezafibrate Infarction Prevention Study Group. Effects of aspirin treatment on survival in non-insulin-dependent diabetic patients with coronary artery disease. *Am J Med*. 1998;105:494-499.
6. Krumholz HM, Chen YT, Radford MJ. Aspirin and the treatment of heart failure in the elderly. *Arch Intern Med*. 2001;161:577-582.
7. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol*. 1999;150:327-333.
8. Lauer MS, Mehta R, Pashkow FJ, Okin PM, Lee K, Marwick TH. Association of chronotropic incompetence with echocardiographic ischemia and prognosis. *J Am Coll Cardiol*. 1998;32:1280-1286.
9. Nishime EO, Cole CR, Blackstone EH, Pashkow FJ, Lauer MS. Heart rate recovery and treadmill exercise score as predictors of mortality in patients referred for exercise ECG. *JAMA*. 2000;284:1392-1398.
10. Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. The fifth report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure (JNC V). *Arch Intern Med*. 1993;153:154-183.
11. Hubbard BL, Gibbons RJ, Lapeyre AC III, Zinsmeister AR, Clements IP. Identification of severe coronary artery disease using simple clinical parameters. *Arch Intern Med*. 1992;152:309-312.
12. Gibbons RJ, Balady GJ, Beasley JW, et al. ACC/AHA Guidelines for Exercise Testing: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Exercise Testing). *J Am Coll Cardiol*. 1997;30:260-311.
13. Snader CE, Marwick TH, Pashkow FJ, Harvey SA, Thomas JD, Lauer MS. Importance of estimated functional capacity as a predictor of all-cause mortality among patients referred for exercise thallium single-photon emission computed tomography: report of 3,400 patients from a single center. *J Am Coll Cardiol*. 1997;30:641-648.
14. Lauer MS, Francis GS, Okin PM, Pashkow FJ, Snader CE, Marwick TH. Impaired chronotropic response to exercise stress testing as a predictor of mortality. *JAMA*. 1999;281:524-529.

How should we stratify on many covariates simultaneously?

- Stratification by Propensity Score Quintile
  - Fit a PS model for each subject
  - Split the subjects into 5 strata (subclasses) of equal size by their propensity scores.
- Five strata of equal size (quintiles) constructed from the PS will usually suffice to remove over 90% of the selection bias due to each of the individual covariates in the PS model. Thm A1 RR'84

Example in Rosenbaum and Rubin 1984

Surgery vs. Medicine for Coronary Artery Disease

- Coronary bypass surgery or medical/drug therapy for coronary artery disease?
  - 1515 subjects – 590 (39%) were surgical patients, the remaining 925 were medical patients.
  - 74 observed covariates describing hemodynamic, angiographic, lab and exercise test results, as well as patient histories and demographics.
  - Each of the 74 covariates was significantly different comparing surgical to medical patients.

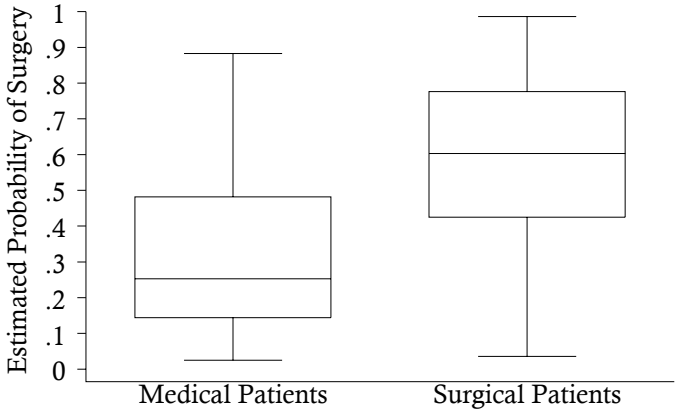
Rosenbaum and Rubin (1984)

Propensity Model for CAD Study

- Logistic regression used to predict treatment assignment for each of the 1515 subjects on the basis of…
  - The 74 covariates themselves
  - Interactions between some covariates
  - Quadratic terms for some covariates
  - Model selection process was sequential – described in the paper…

Rosenbaum and Rubin (1984)

Overlap of Treatment Groups



- For almost every surgical patient, there is a comparable medical patient in terms of having a similar estimated Pr(surgery).

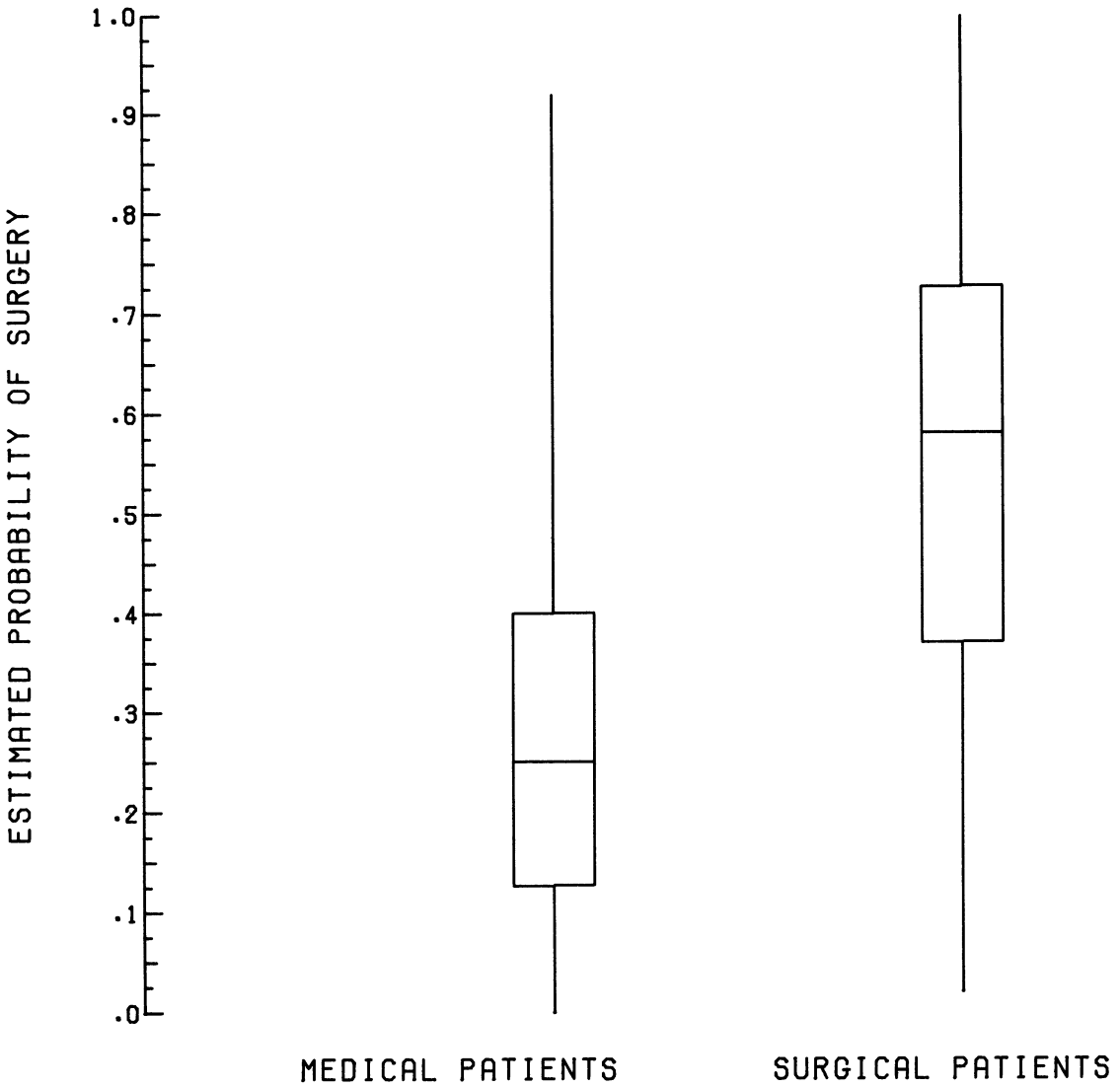


Figure 6. Boxplots of the Estimated Propensity Score.

the late patients. (For proof, see Corollary B.1 of Appendix B.) The observed values of these five covariates were indeed balanced by our procedure: the main-effect F ratios were 2.1, .1, .3, .2, and .0; the interaction F ratios were .4, 1.4, .1, .6, and .3.

3. ESTIMATING THE AVERAGE TREATMENT EFFECT

3.1 Survival; Functional Improvement; Placebo Effects

In this section, we show how balanced subclasses may be used to estimate the average effects of medicine and surgery on survival and functional improvement. Functional capacity is measured by the crude four-category (I = best, II, III, IV = worst) New York Heart Association classification, which measures a patient's ability to perform common tasks without pain. The current study is confined to patients in classes II, III, or IV at the time of cardiac catheterization, that is, patients who could improve. A patient is defined to have *uninterrupted improvement* to t years after cardiac catheterization if he:

1. is alive at t years and
2. has not had a myocardial infarction before t years and
3. is in class I or has improved by two classes (i.e., IV to II) at every follow-up before t years;

otherwise the patient does not have uninterrupted improvement to t years.

It should be noted that there is substantial evidence that patients suffering from coronary artery disease respond to placebos; for a review of this evidence, see Benson and McCallie (1979). Part or all of the difference in functional improvement may reflect differences in the placebo effects of the two treatments.

3.2 Subclass-Specific Estimates; Direct Adjustment

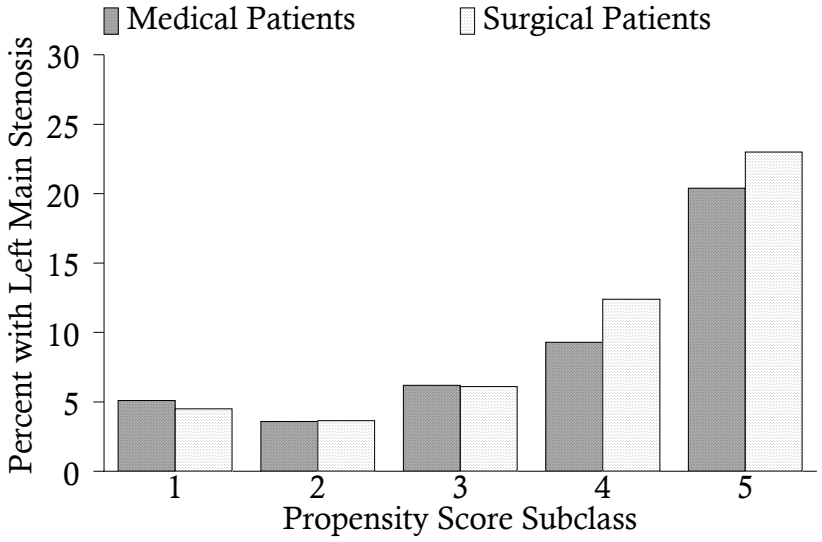
The estimated probabilities of survival and functional improvement at six months in each subclass for medicine and surgery are displayed in Table 1. (These estimates

Propensity Score Subclassification

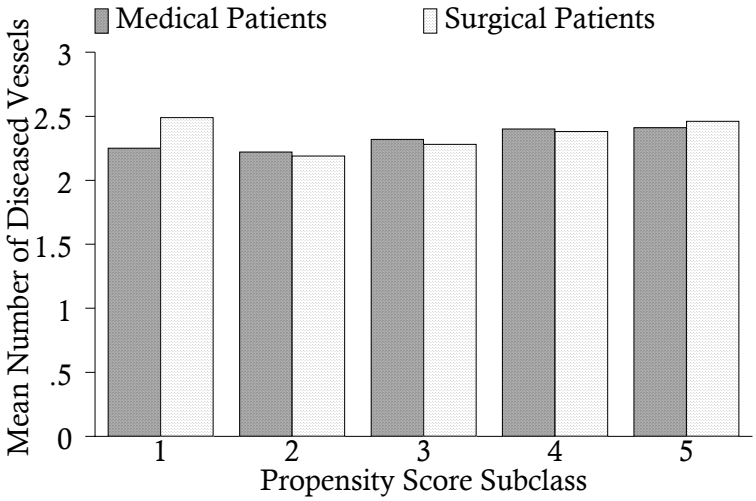
- 1515 patients were divided into five strata of 303 patients each, using estimated propensity scores.

PS Strata	Propensity Score ≈ Prob(surgery | covars.)	Actually got surgery	Actually got medical
5	Highest 303 scores	234 (77%)	69 (23%)
4	2nd highest	164 (54%)	139 (46%)
3	Middle	98 (32%)	205 (68%)
2	2nd lowest	68 (22%)	235 (78%)
1	Lowest 303 scores	26 (9%)	277 (91%)

Balance Within Subclasses: % with Left Main Stenosis



Balance within Subclasses: Number of Diseased Vessels



Subclass Specific Estimates:

Outcomes: Survival at 6 mo & Uninterrupted Improvement at 6 mo

PS Subclass	Group	Patients	P(Survived)	P(Improved)
1	Med	277	.892	.351
	Surg	26	.846	.538
2	Med	235	.953	.402
	Surg	68	.926	.705
3	Med	205	.922	.351
	Surg	98	.898	.699
4	Med	139	.941	.303
	Surg	164	.933	.706
5	Med	69	.924	.390
	Surg	234	.914	.696

An example of where we are going....

Full matching with propensity scores[1]

Ben Hansen

Department of Statistics
University of Michigan, Ann Arbor

Joint Statistical Meetings, August 2005

[1] A presentation (largely) of Hansen, B.B. (2004), Full matching in an observational study of coaching for the SAT, *JASA* **99**, 609–618.

Optmatch creator

Email Address
734-647-5456

Ben Hansen

Research Affiliate, Population Studies Center;
Assistant Professor, Statistics Department;
Faculty Associate, Survey Research Center

Ph.D., University of California, Berkeley
M.A., University of California, Berkeley

Ben Hansen's research interests include optimal matching, propensity-score adjustments for observational studies, quasiexperimental methods, and program assessment. In recent work, he investigates informed consent and perception of risk in survey participation; how to reduce disclosure risk; and how to increase security in the dissemination of human subjects data.

Funded Research:

Human Subjects Protection and Disclosure Risk Analysis (NICHD)

New Publications

Rogowski, Freedman, Schoeni. "Neighborhoods and Health of Elderly." PSC Research Report 06-600.

Geronimus, Hicken, Keene, & Bound. "Age Patterns of Allostatic Load Scores among Blacks and Whites." *AJPH*, 2006.

Farley and Haaga, eds. *The American People: Census 2000.*

Recent Publications

Journal Articles

Evans, S.E., Ben Hansen, P.B. Stark. "Minimax expected measure confidence sets for restricted location parameters." *Bernoulli,* 11:571-590. 2005.

Hansen, Ben. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association,* 99:609-618. 2004.

Full matching with propensity scores. . .

IPTW

- relieves the analyst of the need to reject lots of control subjects in order to get comparable groups;
- can be accomplished with the help of my add-on package for R, optmatch;
- does not ward off problems due to lurking variables, *a.k.a. hidden bias*, or *unmeasured confounding*; but —
- in the absence of hidden bias, should reconstruct a "lurking experiment"; and
- offers greater promise of success at this than either multiple regression or matching with a xed number k of controls.

Oh, did I mention that there is a paper? Hansen, B.B. (2004), Full matching in an observational study of coaching for the SAT, *JASA* **99**, 609–618.

An observational study of effects of coaching for the SAT

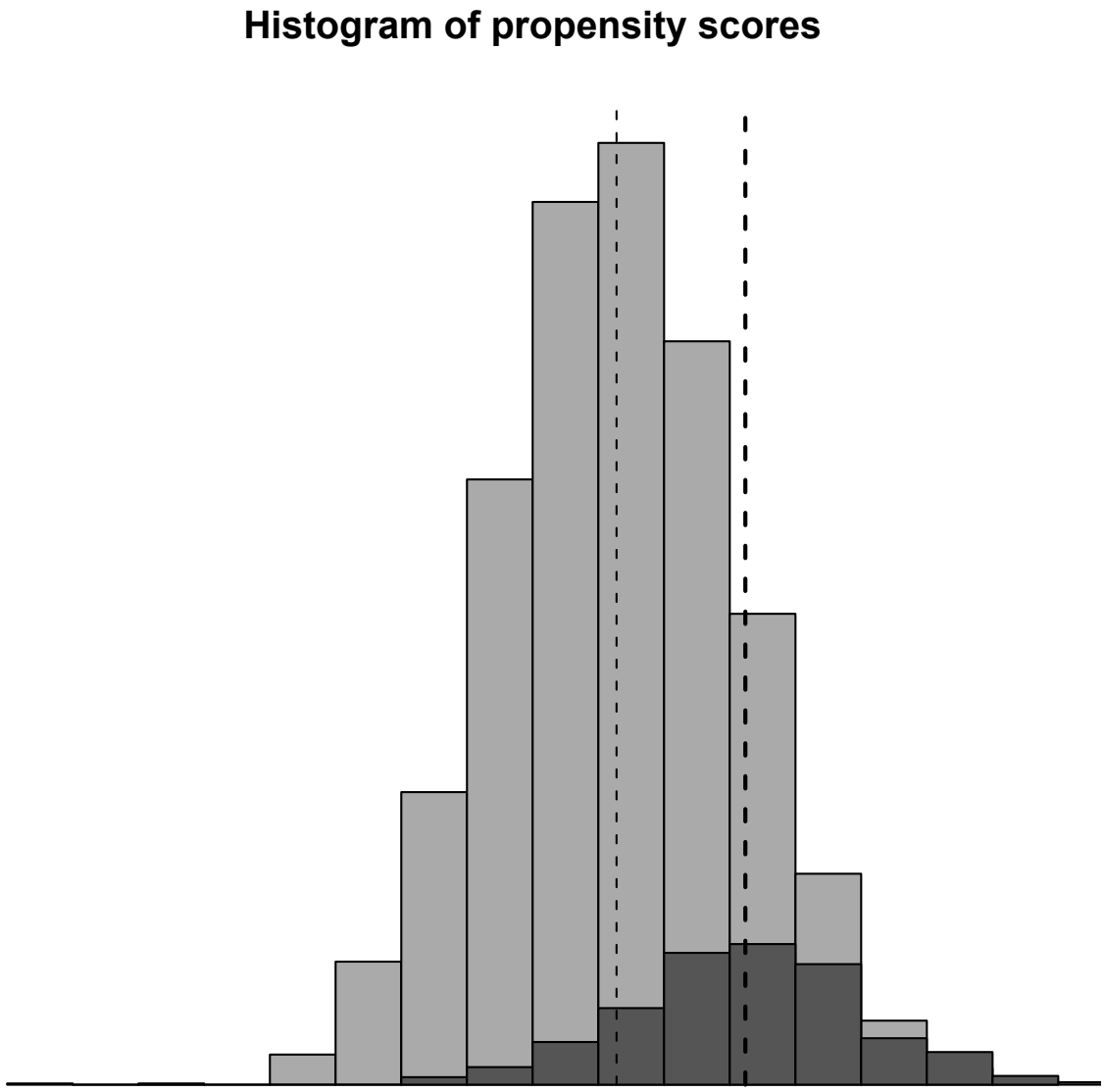
Powers and Rock (1998) sampled one in 200 SAT-I registrants in 1995-96.

- The "treatment" is being coached for the SAT. This information comes from survey responses.
- Outcomes, *i.e.* SAT scores, come from the College Board's administrative records.
- Many students took the SAT or PSAT before being coached; so there are pretest scores too.
- P& R's several analyses tell roughly the same story about coaching and SAT scores.

Challenges confronting analysts of the study

- Ostensibly dissimilar treatment and control groups.
- Up to 40% item non-response among survey respondents.
- Because of item non-response and group dissimilarity, P&R's propensity-matched analysis uses only about 500 of 4000 available observations.
- Their one analysis that dispensed with the fewest observations gave the largest coaching effects. (Coincidence?)

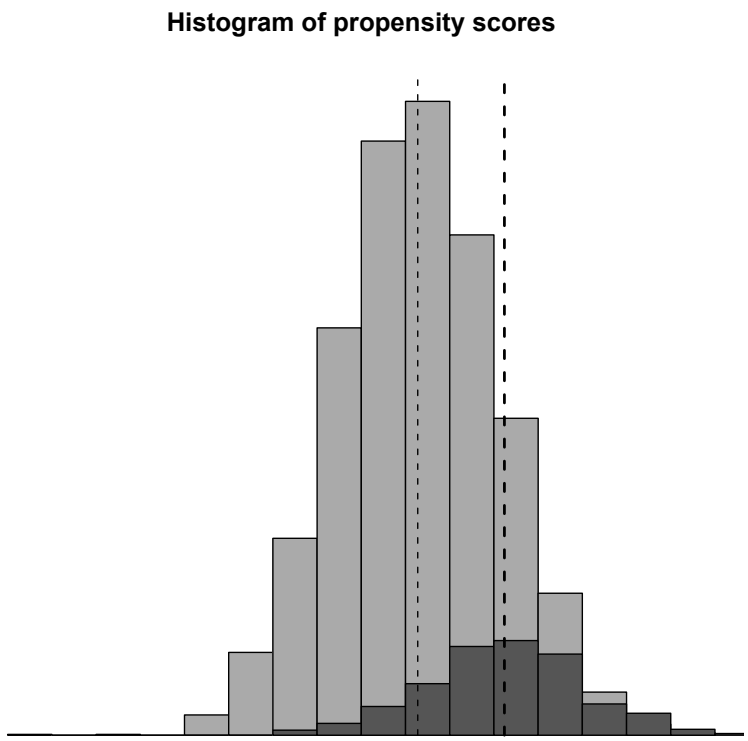
The challenge of matching on the propensity to be coached



Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- Propensity score refers to $\phi(X) = \mathbf{E}(Z|X)$
- $\ldots$ or to $\hat{\phi}(X)$.
- Propensity score$\approx$linear discriminant.

This is typical:



Among matching techniques, only full matching fully adapts...

Fullmatch

Table 1. Selected Pretreatment Variables

Variable	Range of values	Standardized bias	Percentage of sample
Math section of PSAT	20–43	−.1	18
	45–51	.1	17
	52–57	−.1	16
	58–80	.1	15
	Not taken	.1	34
Mean SAT at respondent's first-choice college	787–987	−.3	16
	988–1,060	−.2	16
	1,061–1,123	.1	16
	1,124–1,336	.3	16
	No response	.0	36
Father's education	High school	−.4	40
	A.A. or B.A.	−.1	26
	Graduate	.4	25
	No response	.2	9
Average math grade	"Excellent"	.1	35
	"Good"–"fail"	−.1	59
	No response	.1	6
Foreign language years taken	0–2	−.3	64
	3–4	.3	27
	No response	.1	9

well as scores on previous SAT–I or PSAT tests and their answers to the Student Descriptive Questionnaire (SDQ), which all SAT–I registrants are asked to complete. By their responses to questions about extracurricular SAT preparation, respondents split into a treated and a control group, and the data describe the results of a classical quasiexperiment (Campbell and Stanley 1966).

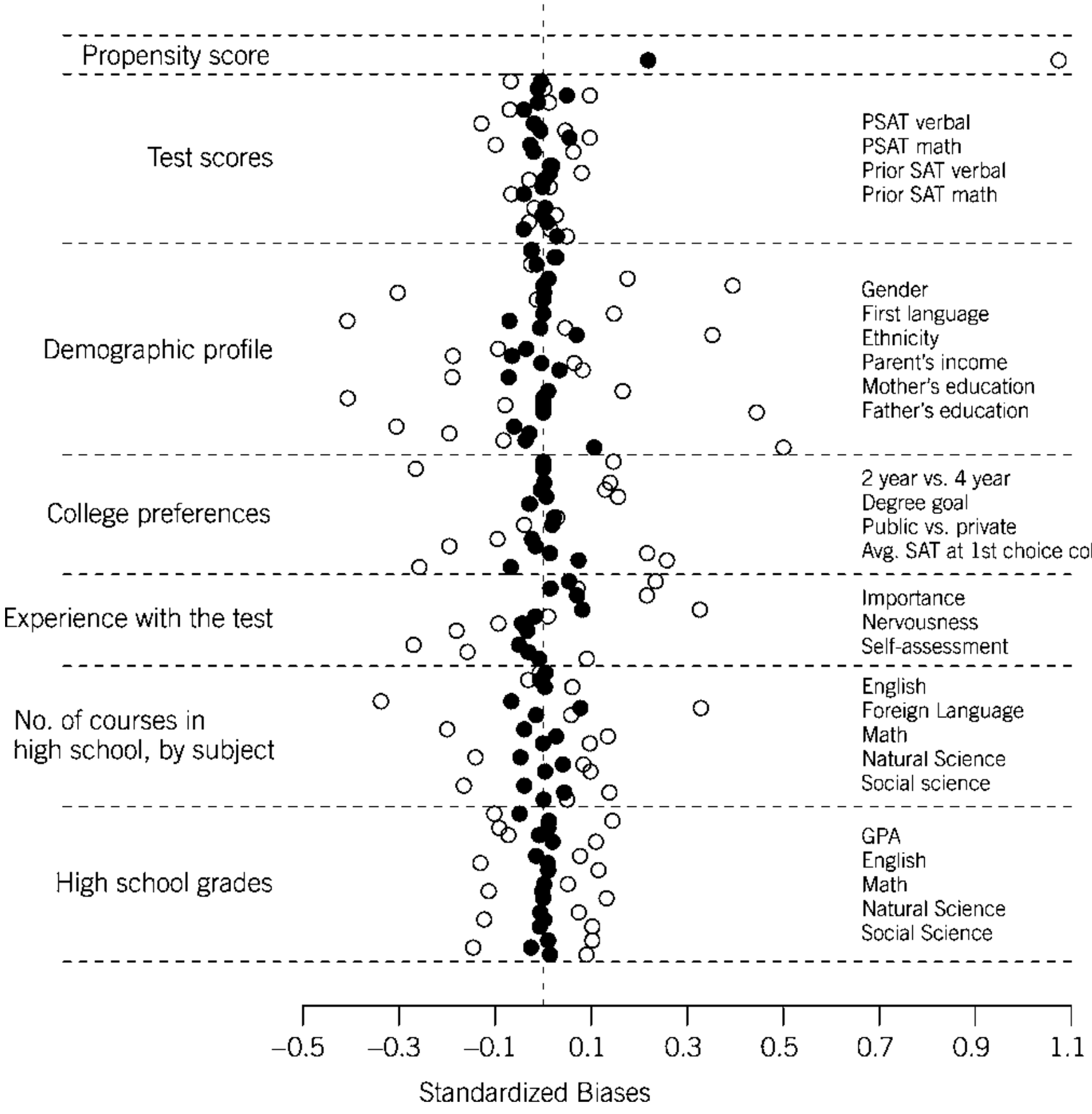
Nineteen in twenty of the survey respondents actually took the spring 1996 or fall 1995 exam for which they had registered. The analysis given below restricts itself to these 3,994 students, using the corresponding SAT scores as outcome measures. Thus the record gives coaching status and SAT outcomes for all students in the sample to be analyzed; among the additional measures, each available for some fraction of the students, are pretest scores, racial and socioeconomic indicators, various data about their academic preparation, and responses to a survey item that, by eliciting students' first choices in colleges, recovered an unusually discriminating measure of students' educational aspirations. In all, there are 27 pretreatment variables.

The coached and uncoached groups differ appreciably in these recorded measures—as do high and low scorers on the SAT. Table 1 offers some illustration of this, giving overall incidences of various covariate attributes and comparing their relative incidences in the coached and uncoached groups. (The statistic here used to effect these comparisons is the *standardized bias*, given for a variable v by $(\bar{v}_t - \bar{v}_c)/s_p$, where $\bar{v}_t$ and $\bar{v}_c$ are the average values of v in the treatment and control groups, respectively, and s_p^2 is the pooled within-group variance in v.) Yet the table shows only five covariates; the analysis must address biases on all 27 of them.

1.2 Missing and Misleading Data in Regression and in Subclassification

In regression-based adjustment, the simplest way to handle missing data on a covariate is to reject cases without complete information. In adjustment based on matching or stratification,

see RItools, PSAgraphics



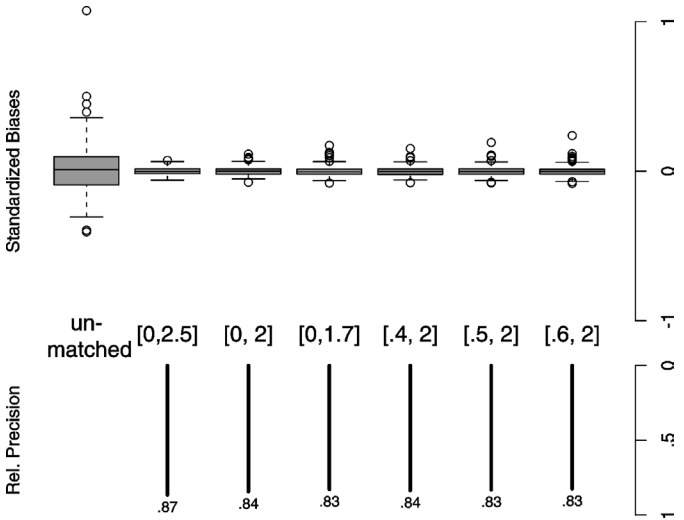


Figure 4. Standardized Biases and Relative Precision [R(·, unconstrained full match)] of Optimal Stratifications With Variously Constrained Match Ratios.

treatment status. These comparisons lead us to prefer a thinning cap of .5. (Had we selected first a thinning and then a thickening cap rather than the reverse, this procedure would have led

the configuration of the matched set into which it is placed, it should be no surprise that the two full matchings lead to similar estimates of the coaching effect.

4. ESTIMATING TREATMENT EFFECTS

To estimate treatment effects, a model such as (1) must be supplemented with a causal formalism and appropriate causal assumptions. For this analysis, the most natural setup is that of Rubin (1977), who posits random variables Y_t and Y_c both for outcomes under the control condition and for outcomes under the treatment condition. Adding the assumption that these variables are conditionally independent of the treatment assignment variable (Z) given the covariates ($\mathbf{X}$) makes inference about treatment effects possible.

Using ETT weighting to combine by-stratum treatment control differences, the [.5, 2] matching leads to aggregate contrasts of 26 points on the math section and 1 point on the verbal. Under causal assumptions as presently discussed, these estimate effects of coaching on the coached. Using model (1), the accompanying standard errors are 5 and 5 points. By contrast, the unadjusted differences of treated and control group means were 41(±5) and 9(±5) points.

As one might expect, those matchings that fail to reduce discernible biases to an indiscernible level give higher effect

breastfeeding

By Nadia Kounang

Updated 4:15 AM ET, Mon March 27, 2017

Women breastfeed their babies at the Hirshhorn Museum in Washington in 2011.

Story highlights

Study finds some short-term cognitive benefit to breastfeeding

Differences between breastfed and non-breastfed children lost by age five

(CNN) — While the medical benefits of breastfeeding for helping newborns fight infections and helping pre-term infants get stronger are fairly well established, the long-term impact is much less so.

While new mothers may debate what they believe to be long-term benefits, a new study published in the journal Pediatrics finds that breastfeeding has little impact on long-term cognitive development and behavior.

The study followed 7,478 Irish children born full term, from the time they were 9 months old. They were then evaluated at three years and again at five years of age.

At three, the children's parents were asked to fill out questionnaires evaluating vocabulary and problem-solving skills to assess cognition and behavior. At age five, both parents and teachers were asked the same questions.

While the researchers found that those children who were
s of
s at three,

Related Article: If 'breast is best' for newborns, where's the support in hospitals?

Short-term benefits

"I think [the study] fits well in the body of literature that long-term benefits of breastfeeding look a whole lot smaller or non-existent if you properly control for your confounding variables," said Dr. Brooke Orosz, a professor of mathematics at Essex County College and adviser to the organization Fed is Best. Orosz was not involved with the study.

Orosz cited that while the study did not take into account maternal IQ, it was able to consider and control for education level and income, which were good proxies.

Considering other factors

Like many other breastfeeding studies, long-term benefits have been associated with breastfeeding, but once socio-economic factors such as education and income are accounted for, the differences between those children who were breastfed and those who weren't are negligible.

Related Article: Extended breastfeeding linked to higher IQ and income in study

"The easy question -- do kids who are breastfed have better outcomes? The answer is yes. The difficult question is: is it breast milk that improves their brain or is it that growing up with parents who are better educated and have better incomes makes a difference?"

While for many new mothers there may be a debate about whether to breastfeed or not, Nancy Hurst, director of Women's Support Services at Texas Children's Hospital Pavilion for Women said that while breastfeeding has many benefits, what was key was nurturing a relationship between mother and child.

"You need to just enjoy the relationship -- that is most important to nurture the mother-baby relationship. Even if at times that doesn't mean exclusive breastfeeding," said Hurst. Hurst is also an international board certified lactation consultant.

Orosz said that for many soon-to-be-newbie parents, it's important to read the literature and really understand what is being evaluated when it comes to breastfeeding.

"Parents need to hear that a lot of it is confounding variables before they are in that situation -- when they are able to process it rationally," she said and added "I think it shouldn't be a debate."

And for those parents who are in the thick of just having had a newborn, hopefully this advice will be heard.

Related Article: 'Their moment, their spac

Breastfeeding, Cognitive and Noncognitive Development in Early Childhood: A Population Study

Lisa-Christine Girard, PhD,[a,b] Orla Doyle, PhD,[b,c] Richard E. Tremblay, PhD[a,b,d,e,f]

abstract

BACKGROUND AND OBJECTIVES: There is mixed evidence from correlational studies that breastfeeding impacts children's development. Propensity score matching with large samples can be an effective tool to remove potential bias from observed confounders in correlational studies. The aim of this study was to investigate the impact of breastfeeding on children's cognitive and noncognitive development at 3 and 5 years of age.

METHODS: Participants included ~8000 families from the Growing Up in Ireland longitudinal infant cohort, who were identified from the Child Benefit Register and randomly selected to participate. Parent and teacher reports and standardized assessments were used to collect information on children's problem behaviors, expressive vocabulary, and cognitive abilities at age 3 and 5 years. Breastfeeding information was collected via maternal report. Propensity score matching was used to compare the average treatment effects on those who were breastfed.

RESULTS: Before matching, breastfeeding was associated with better development on almost every outcome. After matching and adjustment for multiple testing, only 1 of the 13 outcomes remained statistically significant: children's hyperactivity (difference score, −0.84; 95% confidence interval, −1.33 to −0.35) at age 3 years for children who were breastfed for at least 6 months. No statistically significant differences were observed postmatching on any outcome at age 5 years.

CONCLUSIONS: Although 1 positive benefit of breastfeeding was found by using propensity score matching, the effect size was modest in practical terms. No support was found for statistically significant gains at age 5 years, suggesting that the earlier observed benefit from breastfeeding may not be maintained once children enter school.

FREE

[a]*School of Public Health, Physiotherapy, and Sports Science,* [b]*Geary Institute for Public Policy, and* [c]*UCD School of Economics, University College Dublin, Dublin, Ireland; and* [d]*Research Unit on Children's Psychosocial Maladjustment (GRIP), Departments of* [e]*Pediatrics, and* [f]*Psychology, Université de Montreal, Montreal, Canada*

Dr Girard conceptualized the study, carried out the initial analyses, interpreted the data, and drafted the initial manuscript; Drs Doyle and Tremblay conceptualized the study and critically reviewed and revised the manuscript; and all authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

An earlier partial version of this work (age 3 data only) was presented as an oral presentation at the Growing Up in Ireland annual research conference; December 2015; Dublin, Ireland; and at 2 university seminar series; Life Course Centre, University of Queensland, Brisbane, Queensland, Australia and Melbourne Institute of Applied Economic and Social Research, Melbourne, Victoria, Australia; February 2016.

DOI: 10.1542/peds.2016-1848

Accepted for publication Jan 17, 2017

WHAT'S KNOWN ON THIS SUBJECT: The medical benefits of breastfeeding for mother and child are considered numerous, yet the effect of breastfeeding on cognitive abilities remains largely debated given selection into breastfeeding. The effect on behavior is even less well understood.

WHAT THIS STUDY ADDS: In applying quasi-experimental techniques which mimic random assignment, this study supports limited positive impacts of breastfeeding for children's cognitive and noncognitive development. Although significant, the effect of breastfeeding on noncognitive development is small in practical terms.

To cite: Girard L, Doyle O, Tremblay RE. Breastfeeding, Cognitive and Noncognitive Development in Early Childhood: A Population Study. *Pediatrics.* 2017;139(4):e20161848

The medical benefits of breastfeeding for both mother and child are considered numerous and well documented.[1–5] Yet the effect of breastfeeding on general cognitive abilities has been a topic of debate for nearly a century.[6] The mechanism argued to be responsible for these effects is the nutrients found in breast milk.[7,8] Two specific types of long-chain polyunsaturated fatty acids, namely docosahexaenoic (DHA) and arachidonic acid, have been implicated in both visual and neural development and functioning through neural maturation, which is important for cognitive abilities, such as problem solving.[9–11]

The link with nutrients may also impact specific cognitive abilities like language development. For example, language abilities, such as vocabulary, are highly dependent on working and long-term memory given the consolidation and retrieval processes needed during acquisition.[12,13] In rats, deficiency of fatty acids, such as DHA, during lactation resulted in poor memory retention during learning tasks, whereas supplementation of DHA had reversal effects.[14] If the hypothesized "causal" mechanism of superior nutrition in breast milk is true, coupled with the specific impact of DHA on memory, breastfeeding should also impact language abilities. To date, ~20 studies have investigated this association and all but 1[15] examined a combined measure of language (receptive and expressive) or receptive language only. There remains debate as to whether expressive and receptive language in early childhood form distinct modalities of language,[16,17] raising the question of whether breastfeeding would be equally beneficial to each modality in the case of a 2-factor language model.

Less studied is the impact of breastfeeding on behavior. Breastfeeding may lead to reduced behavioral problems as a result of early skin-to-skin contact, which helps form a secure mother-infant bond.[18] Any effects of breastfeeding on cognitive and language development could also prevent the development of behavior problems. The absence of early behavior problems has social, economic, and medical value to society through reduced prevalence of delinquency, incarceration rates, and substance abuse,[19–21] making this an important area of research. With few exceptions, there remains a dearth of high-quality studies examining behavior,[22–25] and among them, consensus is not evident.

Without randomization of mothers to breastfeeding and formula conditions, it is challenging to confirm the causal impact of these hypotheses. One study randomized the provision of a breastfeeding intervention, modeled on the Baby-Friendly Hospital Initiative, and found that the children of mothers in the intervention group had higher intelligence scores compared with controls at age 6 years.[26] The strongest effects were for verbal intelligence. This study offers the best support to date for a causal link between breastfeeding and cognitive development. However, it is the only cluster randomized trial on human lactation.

The majority of studies in this field are observational, thus the causal implications of breastfeeding are questionable given the inherent difficulty in controlling for selection into breastfeeding. For example, initial associations with cognitive development are often reduced after adjustments for confounders, such as parental education/IQ (ie, from an average 5-point to 3-point difference[27]), and, in some cases, the associations are no longer statistically significant.[28] A variety of observational studies now apply quasi-experimental methods to better address the issue of selection bias, making inroads toward a better understanding of potential causal paths. The techniques used include propensity score matching (PSM), instrument variables, and sibling pair models. This study uses PSM because the sibling pair model limits the available pool of participants and instrument variables are extremely sensitive to the validity of the chosen instrumentation, which should be associated with the exposure but not with the outcome except for via the exposure.

Using a large longitudinal population sample, we applied PSM, which mimics random assignment, in an effort to investigate the potential impacts of breastfeeding on children's cognitive ability, expressive vocabulary, and behavior problems. Both breastfeeding duration and intensity were examined. Significant advantages for children who were breastfed, after matching, were expected for all outcomes. Grounded in the recommendations of the World Health Organization,[29] it was expected that larger effect sizes would be observed for children who were fully breastfed and for longer durations.

METHODS

Participants

Participants included families enrolled in the Growing Up in Ireland infant cohort. Families with infants born between December 2007 and May 2008 were identified from the Child Benefit Register and randomly selected to participate. The overall recruitment response rate was 65% (N = 11 134). A detailed description of the study design can be found elsewhere.[30] We used data collected at 9 months and 3 and 5 years of age. Only families with complete data for all confounders when children were 9 months and children who were born full term were included (N = 9854; 88.5% of the initial sample). Boys represented 50.6% (N = 4991)

TABLE 1 Family, Maternal, Infant, and Medical Characteristics: Infant Cohort at 9 Months

	Ever Breastfed ($N = 5940$)	Never Breastfed ($N = 3914$)	*P*
	n (%)	*n* (%)	
Resident spouse/partner (yes)	5469 (92.1)	3213 (82.1)	≤.001
Social class			≤.001
Professional/managerial	3486 (58.7)	1449 (37.0)	
Nonmanual/skilled manual	1533 (25.8)	1419 (36.3)	
Semiskilled/unskilled	505 (8.5)	397 (10.1)	
Unknown/never worked	416 (7.0)	649 (16.6)	
Medical card status (yes)	1336 (22.8)	1433 (36.6)	≤.001
Maternal education			≤.001
Primary level/no education	65 (1.1)	152 (3.9)	
Second level	1782 (30.0)	2269 (58.0)	
Third level	4093 (68.9)	1493 (38.1)	
Maternal working status (yes)	4828 (81.3)	2865 (73.2)	≤.001
Maternal age, y			≤.001
≤ 24	456 (7.7)	653 (16.7)	
25–29	1178 (19.8)	883 (22.6)	
30–34 y	2202 (37.1)	1240 (31.7)	
≥35 y	2104 (35.4)	1138 (29.1)	
Maternal ethnicity (Irish)	4209 (70.9)	3725 (95.2)	≤.001
Maternal depression (yes)	222 (3.7)	201 (5.3)	.001
Smoking in dwelling during pregnancy (yes)	1535 (25.8)	1646 (42.1)	≤.001
Delivery mode (cesarean)	1348 (22.7)	1063 (27.2)	≤.001
Birth weight (≥2500 g; yes)	5842 (98.4)	3810 (97.3)	≤.001
Visit to the NICU (yes)	575 (9.7)	420 (10.7)	.090
Infant sex (boy)	2944 (49.6)	2047 (52.3)	.008
Siblings living in dwelling (yes)	3248 (54.7)	2614 (66.8)	≤.001

Medical card coverage is a means tested card issued by health services on the basis of financial need. There are 2 tiers of medical card coverage: "full coverage," which includes visits to general practitioners plus prescriptions and "general practitioner only coverage," which excludes prescriptions. Regarding the maternal education variable, primary level/no formal education is approximately equivalent to having an elementary to middle school education in the US system; second level is approximately equivalent to a high school diploma or technical trade/vocational diploma in the US system; and third level is equivalent to a college or bachelor's degree, graduate degree, or doctorate. Maternal working status refers to employment before pregnancy. Categorization of maternal depression refers to a score of ≥11 on the Center for Epidemiologic Studies Depression Scale.

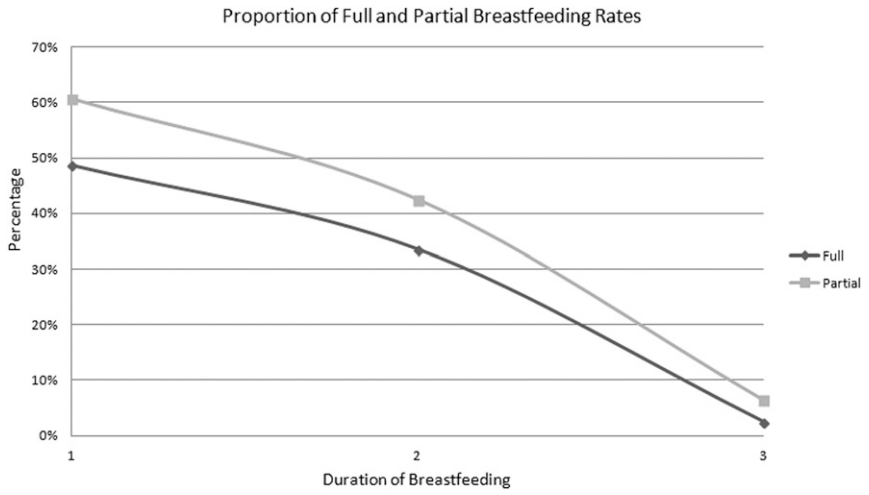


FIGURE 1
The category "1" on the x-axis represents breastfeeding up to 31 days; "2" represents between 32 and 180 days; and "3" represents ≥181 days.

of the sample. Attrition across waves reduced the sample size to 8715 children at 3 years and 8032 at 5 years. Some children had missing data on the cognitive and vocabulary scales, resulting in 8535 and 8241 children respectively at age 3 and 7972 and 7942 children respectively at age 5. Additionally, missing teacher reports for behavior at age 5 years resulted in 7478 children being included in these analyses. Demographic characteristics of the families and rates of breastfeeding engagement can be found in Table 1 and Fig 1. Ethics approval was obtained from the Research Ethics Committee, Department of Children and Youth Affairs Ireland, and written consent was collected from parents/guardians before data collection.

Measures

Children's cognitive abilities and expressive vocabulary were measured by using 2 scales from the British Abilities Scale[31]. The pictures similarities scale assessed problem-solving skills and the naming vocabulary scale assessed expressive vocabulary. The construct validity of each scale was derived by using the Wechsler Preschool and Primary Scale of Intelligence-Revised

TABLE 2 Bivariate Correlations Between Parent and Teacher SDQ Scores and Means (SDs) of Children's Outcomes at 3 and 5 Years of Age

	Conduct Problems, 5 y (Teacher)	Hyperactivity, 5 y (Teacher)	Difficulties, 5 y (Teacher)	Means (SD)	Minimum–Maximum
Conduct problems, 5 y (parent)	$r = 0.23^{***}$	$r = 0.21^{***}$	$r = 0.22^{***}$	1.44 (1.46)	0–10
Hyperactivity, 5 y (parent)	$r = 0.22^{***}$	$r = 0.35^{***}$	$r = 0.32^{***}$	3.23 (2.40)	0–10
Difficulties, 5 y (parent)	$r = 0.22^{***}$	$r = 0.29^{***}$	$r = 0.32^{***}$	7.10 (4.71)	0–32
Conduct problems, 5 y (teacher)	—	—	—	0.73 (1.33)	0–10
Hyperactivity, 5 y (teacher)	$r = 0.51^{***}$	—	—	2.96 (2.81)	0–10
Difficulties, 5 y (teacher)	$r = 0.70^{***}$	$r = 0.82^{***}$	—	5.92 (5.25)	0–32
Conduct problems, 3 y (parent)	—	—	—	2.15 (1.80)	0–10
Hyperactivity, 3 y (parent)	—	—	—	3.10 (2.14)	0–10
Difficulties, 3 y (parent)	—	—	—	7.71 (4.53)	0–32
Nonverbal reasoning, 5 y	—	—	—	58.89 (10.61)	20–80
Nonverbal reasoning, 3 y	—	—	—	53.30 (10.77)	20–80
Expressive vocabulary, 5 y	—	—	—	55.27 (12.22)	20–80
Expressive vocabulary, 3 y	—	—	—	51.16 (12.75)	20–80

*** $P \leq .001$.

(r = 0.74 and 0.83, respectively).[31] Standardized scores that adjusted for performance as compared with other children of the same age, with a mean of 50 and a SD of 10, were used. Age was adjusted in 3-month age bands.

The Strengths and Difficulties Questionnaire (SDQ[32]) was used to assess children's problem behaviors. The parent version was used at age 3 years and both the parent and teacher versions were used at age 5 years. The SDQ is comprised of 5 scales (emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behavior) with ratings of applicability of behaviors on a 3-point scale. A total difficulties scale is included, combining the 4 problem scales, to yield an overall difficulties score. We used the conduct problems, hyperactivity/inattention, and difficulties scales given our focus on externalizing problems. Validation of the SDQ has been extensively documented.[33] Table 2 reports the correlations between parent and teacher SDQ reports and the means and SDs for all child outcomes.

Breastfeeding information was collected retrospectively when infants were 9 months old via maternal report. Support for the reliability of recall in previous breastfeeding studies has been established.[34] However, given the lower reliability regarding the timing of the introduction of additional fluids/solids, Labbok and Krasovec's definition of full (ie, exclusive or almost exclusive) and partial breastfeeding are used.[35] Two breastfeeding variables were created to assess whether the infant was fully or partially breastfed and the duration of each. Mothers were asked 4 questions: "Was <baby> ever breastfed," "How old was <baby> when he/she completely stopped being breastfed," "Was <baby> ever exclusively breastfed," and "How old was <baby> when he/she completely stopped being exclusively breastfed?" First, infants were grouped by breastfeeding status, both full and partial (5940) and never breastfed (3914). Of those who had ever been breastfed, 4795 had full breastfeeding at some point. Next, breastfeeding duration was grouped into 3 intervals; breastfed up to 31 days, 32 to 180 days, and ≥181 days. Each category of duration was treated as mutually exclusive, dummy coded, and compared against infants who had never been breastfed for the purpose of matching.

Confounders have been suggested in part to account for the associations found between breastfeeding and child outcomes. We matched groups (breastfed, never breastfed) on 14 of the most pertinent factors. At the child level, factors included sex (boy/girl), birth weight (≥2500 g), and having neonatal intensive care (yes/no). At the maternal level, factors included age (≤24 years, 25–29 years, 30–34 years, or ≥35 years), highest level of education (primary level/no education, second level, or third level), working status before pregnancy (yes/no), ethnicity (Irish, any other white background, African or any other black background, Asian background, or other, including mixed background), depression (a score of ≥11 on the Center for Epidemiologic Studies Depression Scale), and type of delivery (vaginal or caesarean). Family-level factors included having a partner in the residence (yes/no), social class (professional/managerial, other nonmanual/skilled manual, or semiskilled/unskilled), medical card status (free medical care, free general practitioner care, or no free medical care), total number of household members who smoked during the pregnancy (none, or ≥1), and whether the cohort infant had siblings living in the household.

Statistical Analysis

PSM reduces selection bias by matching children who were breastfed to children who were not, but who had a similar probability of being breastfed based on their measured characteristics. We used PSM logit models with nearest neighbor 1:1 matching techniques. In nearest

neighbor matching, the sample is randomly ordered with matching occurring sequentially between the treatment (breastfed) and control (not breastfed) group based on participants' propensity scores. Typically, the pair is then removed from the list and the next match is created. To ensure optimal matches, we imposed a caliper so that pairs could only be matched if the propensity score was within a tenth of a SD of the other. We also allowed matching with replacement given the low rates of longer durations and full breastfeeding in this cohort. Although matching with replacement has been argued to increase variance in the data, it also arguably reduces bias in the sample by ensuring better quality of matches.[36] Balance checks in all models revealed substantial reductions of bias between matched groups on all individual confounders (ie, 0%–13.9% remaining bias in partial breastfeeding models, 0%–18.1% remaining bias in full models; data available on request). The remaining overall mean bias across models ranged from 3.2% to 8.5%. The ≤20% remaining bias has been suggested as the acceptable cutoff after matching.[37] Thus, we concluded that the analytic matching technique resulted in good matches between conditions. Matching resulted in all participants falling within the area of common support. The average treatment effect on those who were treated (ie, children who were breastfed) is reported. Adjustments were made for multiple hypothesis testing by using the Holmes-Bonferroni method. All statistical analyses for PSM were conducted by using Stata version 13 software (Stata Corp, College Station, TX).

To note, although PSM is advantageous in mimicking random assignment, a drawback is the challenge in evaluating a linear dose-response association, which has previously been found. Structural equation modeling (SEM) offers an alternative approach to examining this dose-response association. Additionally, SEM uses the full sample and has greater power. Thus, the data were also modeled by using SEM, where confounders were treated as correlated exogenous variables, the duration of breastfeeding was treated as a continuous mediating variable, and child outcomes were treated as correlated, which could be influenced by both breastfeeding and confounders. These results can be found in the Supplemental Material.

RESULTS

Postmatching results for children fully breastfed up to 31 days revealed no statistically significant differences between groups on any outcome at age 3 or 5 years (Table 3). Similarly, for children who were fully breastfed between 32 and 180 days, no statistically significant differences were found for any outcomes at either age postmatching (Table 4). Finally, for children who were fully breastfed for ≥6, statistically significant differences were found postmatching for only 2 outcomes, problem solving and hyperactivity at age 3 years. Children who were fully breastfed scored 2.95 (SE = 1.39, $P = .048$) points higher on the problem-solving scale compared with children who were never breastfed and −0.84 (SE = 0.25, $P \leq .001$) points lower on the hyperactivity scale. After adjustment for multiple testing, cognition was no longer statistically significant. However, children who were fully breastfed had slightly lower parent-rated hyperactivity compared with controls, and this remained statistically significant after adjustment (Table 5). Of note, results of the partial breastfeeding models were similar to the full models, however, after adjustment for multiple testing, neither cognitive ability nor hyperactivity at age 3 years remained statistically significant. These results can be found in the Supplemental Material.

DISCUSSION

Without randomized controlled trials, the issue of causality will necessarily remain open, however the present results contribute important insights to the long-standing debate of potential "causal effects" versus artifacts of confounding that are not properly accounted for. This study also provides new perspectives on breastfeeding and children's externalizing behavior. To the best of our knowledge, this is among the first studies to examine expressive vocabulary as an individual outcome and to consider externalizing behavior. It should be noted that our results apply only to infants born full term.

After adjustment for multiple testing, the initial support found for breastfeeding and better problem solving at age 3 years if the child was breastfed for a minimum of 6 months was no longer statistically significant. In addition, no statistically significant effects were found for cognitive ability at age 5 years. These results are in contrast to some studies that have used PSM techniques to examine the effects of breastfeeding and general cognitive abilities.[38–40] However, differences in both analytical choices of the PSM approach used (eg, replacement, calipers) and differing selection of covariates may help to explain these differences across studies. Nonetheless, our findings were surprising in the context of the nutrients in breast milk being responsible for increased cognitive development. Regarding expressive vocabulary, no statistically significant advantages were observed for children who were breastfed at either age 3 or age 5.

The limited research on breastfeeding and behavior problems is inconsistent, despite the relatively consistent reliance on the SDQ. Of interest, studies that have dichotomized the SDQ scales into abnormal scores (ie, at the 85th or 90th percentile) have not found