

Center for  
Health Care  
Research &  
Policy

# SMDM Short Course on Propensity Methods Part Four

**THOMAS E. LOVE, PH. D.**  
**OCTOBER 21, 2001**

Center for  
Health Care  
Research &  
Policy



## Part Four Topics

- A few thoughts on instrumental variables
- How do we know that the propensity score “worked”?
- Addressing your remaining concerns
- What should “always” be done?

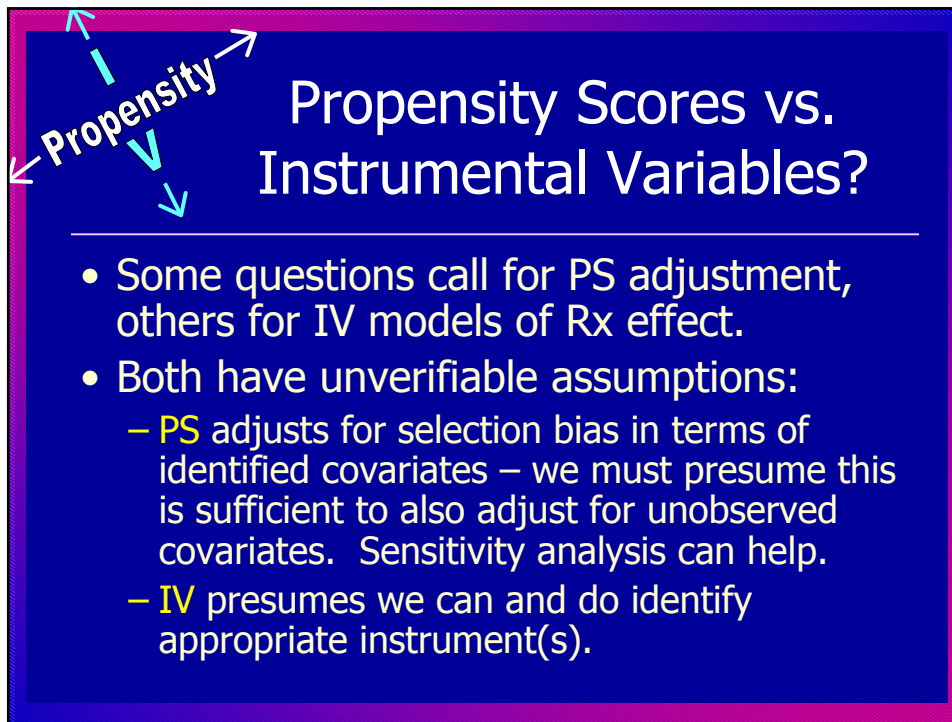
## What About Instrumental Variables?

- **Idea:** Find a variable (the instrument)
  - strongly correlated with the treatment choice
  - but having **no direct effect** on the outcome (outside of the instrument's influence on treatment selection)
- If these two conditions are not met, then IV is not a useful approach.
- In health care, treatment selection is usually closely linked to outcome.

## When Are Instrumental Variables Methods Especially Attractive?

### **An instrument is available, and ...**

- Assignment to a treatment is ignorable, but compliance with the assignment is not perfect so that the **dose** of treatment received is non-ignorable.
- Data are **weak**, in the sense that observed covariates provide insufficient insight into the background to allow estimated effects (adjusting for covariates) to be due to treatment.



## Propensity Scores vs. Instrumental Variables?

- Some questions call for PS adjustment, others for IV models of Rx effect.
- Both have unverifiable assumptions:
  - **PS** adjusts for selection bias in terms of identified covariates – we must presume this is sufficient to also adjust for unobserved covariates. Sensitivity analysis can help.
  - **IV** presumes we can and do identify appropriate instrument(s).

## When is it sufficient to adjust for the **observed** covariates?

- Main Statistical Assumption: Strongly Ignorable Treatment Assignment
- After adjustment for observed covariates, we assume that the different treatment groups are comparable.
- Treatment assignments are unrelated to potential outcomes within strata defined by covariates.

How do we know that the PS  
worked?

## Checking for Covariate Balance

Are we making  
comparisons between  
comparable subjects?

## Checking for Covariate Balance: Large Tabular Presentations

Variable	(Rx) RP %	(Ctrl) RT %	Unadjusted Wald F (p)	Wald F (p) <b>adj. for PS</b>
Incontinent	3	8	12.2 (<.001)	0.09 (.76)
Impotent	21	38	36.1 (<.001)	0.85 (.36)
CHF	5	8	3.8 (.05)	0.20 (.66)
Lung Dx	7	12	5.7 (.02)	0.02 (.89)
Hypertens.	41	45	1.2 (.28)	0 (.98)
Angina	9	18	17.0 (<.001)	0.25 (.62)

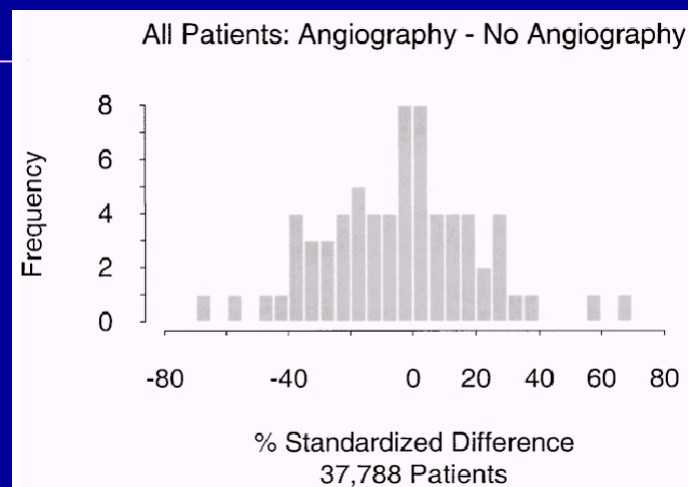
Adapted from Table 1 in Potosky et al. (2000) p. 1585

## Does **Matching** By Propensity Scores Help Reduce Selection Bias?

Standardized Differences are an Appropriate Summary Statistic to Use in Assessing Covariate Balance

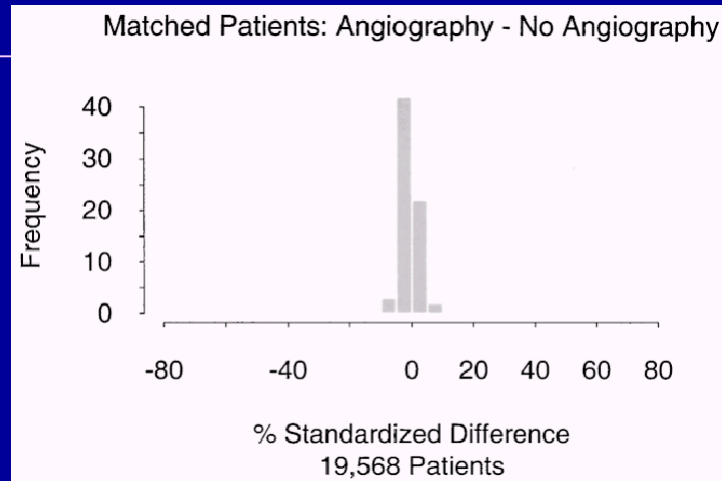
$$d = \frac{100(\bar{x}_{Treatment} - \bar{x}_{Control})}{\sqrt{\frac{(s_{Treatment}^2 + s_{Control}^2)}{2}}}$$

## Standardized Differences (%) in Covariate Means: **Before** Matching



Normand et al. (2001) p. 395

## Standardized Differences (%) in Covariate Means: **After** Matching



Normand et al. (2001) p. 395

## Possible Discussion Issues: Diagnostics & Display

- Modeling the propensity score
  - Handling missing data
  - Assessing the PS model
- Checking and presenting the results of a PS-based analysis
  - In matching
  - In subclassification (stratification on PS)
  - In multivariate adjustment

END

# How Should the Propensity Score Be Modeled?

Handling missing data  
Assessing the model

## Missing Data and Generalized Propensity Scores

- Pattern of missing covariates can be prognostically important
- PS should, ideally, condition both on
  - Observed values of covariates, and
  - Observed missing-data indicators
- Generalized PS lead in expectation to balanced missing data patterns & covariate distributions in the treatment and control groups
- D'Agostino Jr & Rubin (2000) provide details.

Discussion Topics

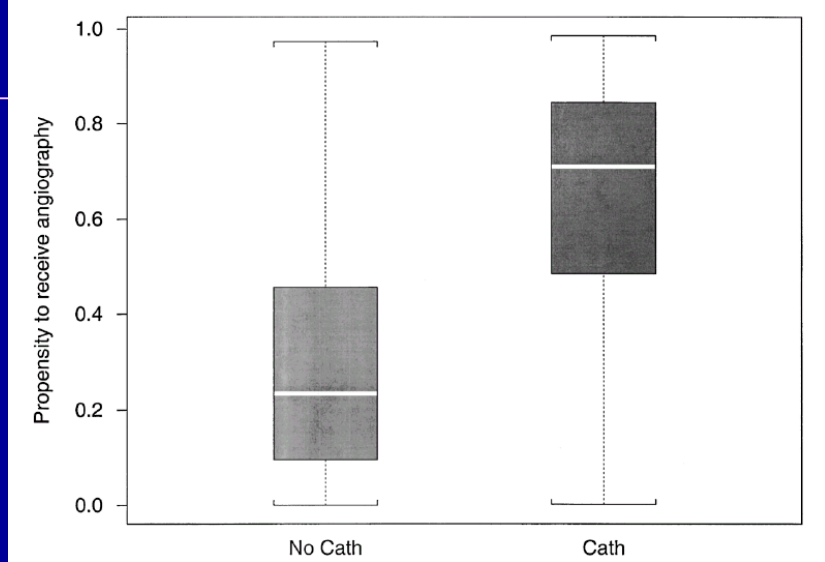
## Dealing with Missing Data: A Typical Approach

- MAR = assumed missing at random – mechanism by which data are missing is unrelated to information not in  $\mathbf{X}$ .
  - **Discrete**: include binary “missing”  $x$
  - **Continuous**: fit two predictors:
    - [1] subject was measured/unmeasured,
    - [2] if subject measured, then value.

## Assessing the Propensity Score Model: Quality of Fit

- The main goal remains covariate balance.
- Do propensity scores accurately track treatment selection?
  - Area under ROC curve
  - Hosmer-Lemeshow Goodness of Fit test
- Do estimated propensity scores (or covariates) overlap by treatment groups?

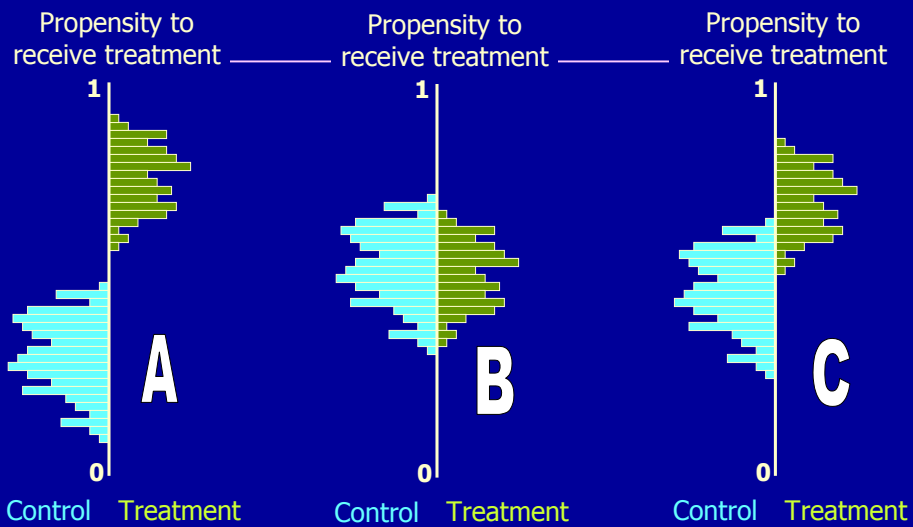
## Do the Propensity Scores Overlap?



Normand et al. (2001) p. 394

### Discussion Topics

## How Much Overlap Do We Want?



## Using Propensity Scores to Match Subjects

- What are the most important concerns?
- What's the tradeoff between inexact and incomplete matching?
- Can simple schemes be effective?
- Did the matching "work"?

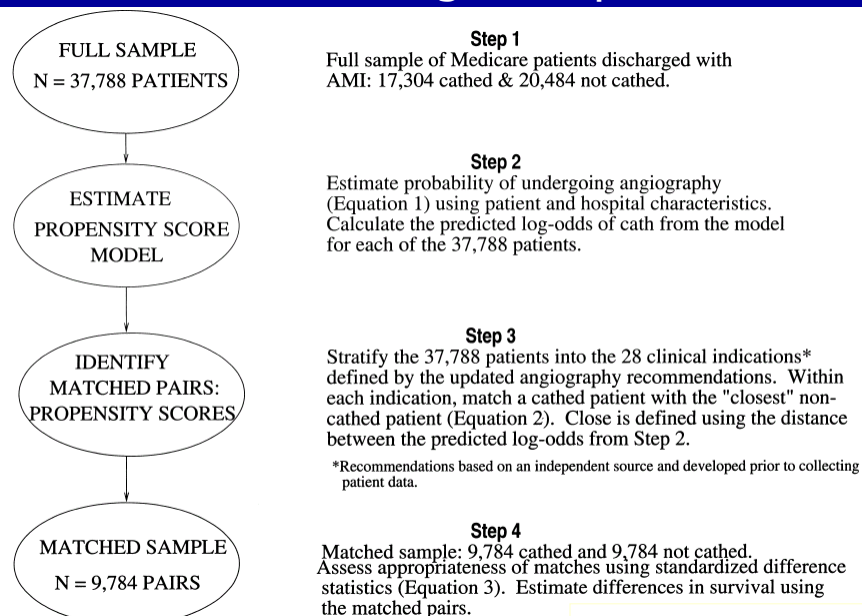
## Incomplete vs. Inexact Matching

- Trade-off between
  - Failing to match all treated subjects (**incomplete**)
  - Matching dissimilar subjects (**inexact** matching)
- In practice, concern has been inexactness
- Severe bias due to incomplete matching – try to **match all treated subjects**, then follow up with analytical adjustments for residual imbalances in the covariates

## Choosing Matches

- Available matching schemes:
  - Match on Mahalanobis distance of a set of important covariates
  - Match directly on propensity score
  - Match on logit of (PS)
  - Match using PS calipers and a distance
- Cutoffs for acceptable matches
- “Nearest Available” vs. Optimal Matching

## A Matching Example



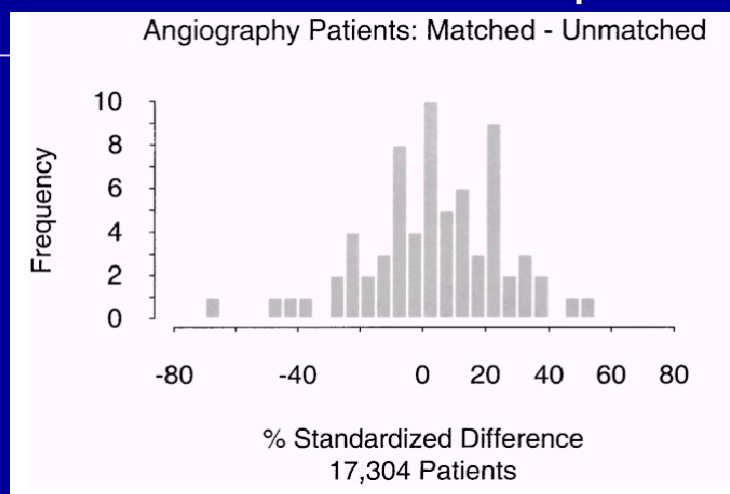
Normand et al. (2001) p. 390

## Example Details: Using the PS To Match in the Angiography Study

- Stratify by clinical indication (28 groups)
- Within each group, used nearest available matching on the estimated logit of PS
- Cutoff: within .60 SE of the logits
- 57% of "catheterized" pts were matched
- Are there important differences between matched and unmatched patients?

## Incomplete Matching: Matched and Unmatched Catheterized patients

Discussion Topics



Normand et al. (2001) p. 395

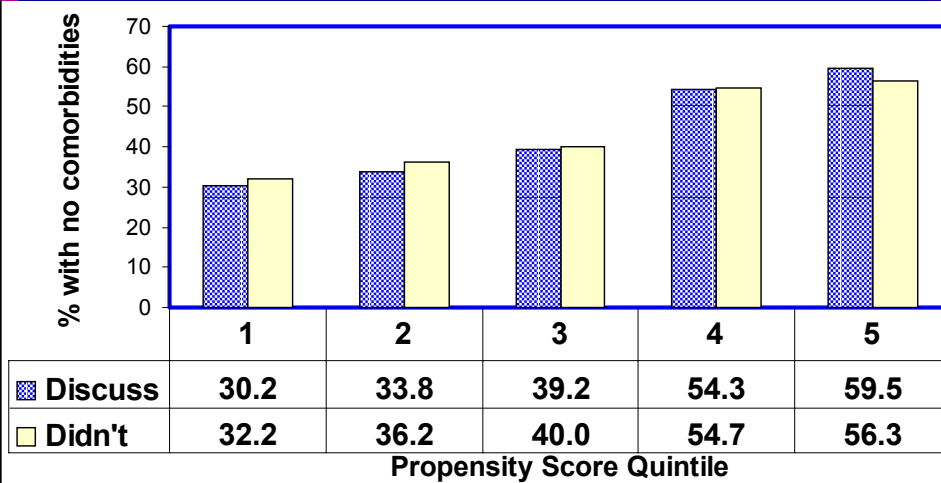
## Using Propensity Scores to Subclassify (Stratify) Subjects

How many subgroups?  
Checking the model  
Combining & Presenting results

## Using Subclassification (Stratification) on the PS

- Five subclasses (quintiles) constructed from the PS will often suffice to remove over 90% of the selection bias due to each of the covariates.
- Decide whether stratum boundaries will be based on PS for both groups combined (typical situation), or in treated or control group alone.
- Once the subclasses are defined, treated and control subjects in the same subclass are compared directly.

## Balance within PS Subclasses: Comorbidities Status (*Simulated*)



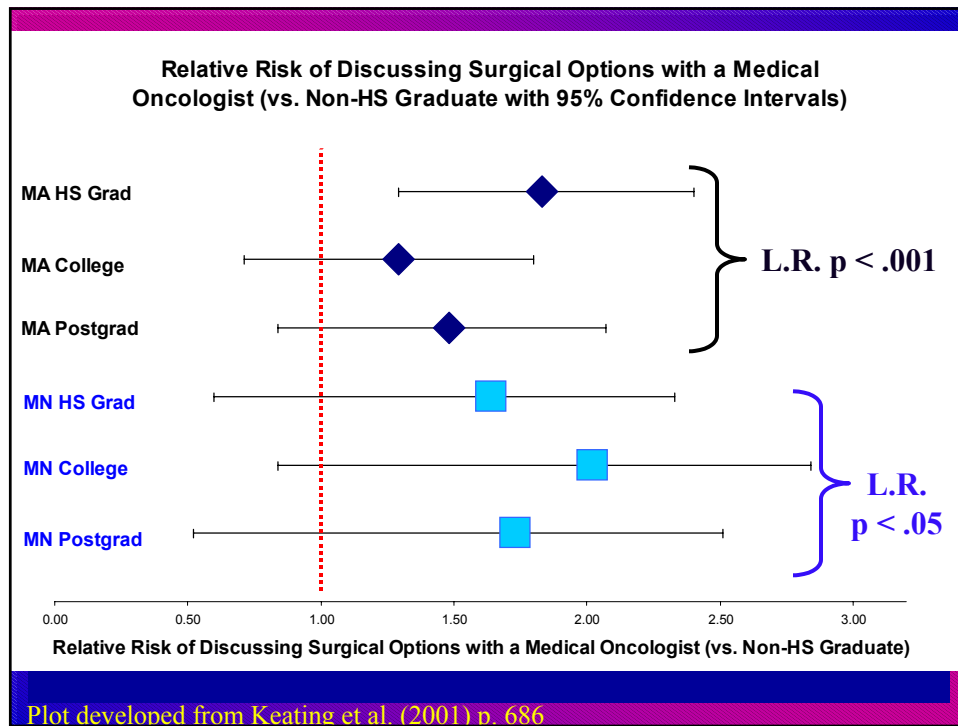
Simulated based on Keating et al. (2001)

## Factors Associated with Discussing Surgical Options with a Medical Oncologist

Characteristic	Mass. RR (95% CI)	Minn. RR (95% CI)
White race (vs. Nonwhite)	1.09 (0.75, 1.41)	1.03 (0.56, 1.66)
HMO insurance	1.50 (1.31, 1.67)	0.98 (0.81, 1.16)

This is a small part of the table – each RR  
is adjusted for 13 covariates in PS model

Adapted from Keating et al. (2001) p. 686



Discussion Topics

## Using the Propensity Score to Subclassify

- Subclassify (stratify) into 5 groups by PS
  - Estimate  $pr(\text{outcome})$  within PS group both for treated & control subjects
  - Combine estimates across groups to get global estimates for treated & control
  - Use global estimates to calculate absolute difference in  $pr(\text{outcome})$ , or other summary measure(s).
  - SAS code provided by D'Agostino (handout)

## Using Propensity Scores in Multivariate Adjustments

How can results be presented  
effectively?

Multivariate Matching vs.  
Multivariate Adjustment

## Reporting Regression Models

**TABLE III** Multivariate Predictors of Inpatient Mortality in Patients <65 Years Old

Variable	OR	95% CI
Care by cardiologist	1.051	0.639-1.593
Propensity score	0.729	0.193-2.697
Age	1.027	1.001-1.054
Caseload (1-6 cases)	1.427	0.844-2.420
Caseload (7-12 cases)	1.194	0.714-1.996
Caseload (13-23 cases)	1.008	0.642-1.580
CABG Hospital	0.543	0.361-0.805
ASG score 2	2.460	1.501-4.190
ASG score 3	9.473	5.678-16.409
ASG score 4	32.912	16.040-68.314
Atrial fibrillation or flutter	1.202	0.645-2.111
Ventricular fibrillation	1.569	0.811-2.886
Cardiogenic shock	12.499	7.912-19.780
Cardiomyopathy	2.062	0.744-4.857
Diabetes mellitus	0.930	0.650-1.318
Dialysis	1.534	0.525-4.160
Women	1.338	0.954-1.866
Heart failure	1.121	0.782-1.591
History of CABG	2.717	1.605-4.462
Chronic renal failure	2.124	0.790-5.152
Acute renal failure	1.899	0.841-4.067
Malignant neoplasm	1.102	0.232-3.718

CABG = coronary artery bypass grafting.

Hosmer-Lemeshow  $\chi^2 = 10.78$ , 8 df (p = .21)

**TABLE V** Multivariate Predictors of Inpatient Mortality in Patients  $\geq 65$  Years Old

Variable	OR	95% CI
Care by cardiologist	0.862	0.721-1.030
Propensity score	0.567	0.326-0.976
Age	1.038	1.028-1.047
Caseload (1-6 cases)	1.135	0.912-1.413
Caseload (7-12 cases)	1.047	0.846-1.296
Caseload (13-23 cases)	1.023	0.839-1.247
ASG score 2	1.555	0.919-2.856
ASG score 3	4.147	2.464-7.591
ASG score 4	18.772	10.894-34.977
Ventricular fibrillation	5.025	3.674-6.861
Cardiogenic shock	13.832	10.958-17.553
Dialysis	1.262	0.718-2.168
Women	0.901	0.791-1.026
Hypertension	0.763	0.579-0.994
History of CABG	1.063	0.785-1.421
Chronic renal failure	1.080	0.752-1.527
Acute renal failure	3.254	2.540-4.165
Neoplasm	0.870	0.572-1.286

Abbreviations as in Table III.

Hosmer-Lemeshow  $\chi^2 = 6.43$ , 8 df (p = .60)

Nash et al. (1999) p. 653

## Often, Multivariate Adjustment leads to large tables & long footnotes

Table 3. Comparison of 24-month survey responders on individual urinary, bowel, and sexual domain items\*

Domain	RP†,‡ (n = 961)	RT‡,§ (n = 373)	Odds ratio (95% confidence interval)
<b>Urinary</b>			
No control or frequently leaks or drips urine vs. total control or occasionally leaks	9.6 (9.8)	3.5 (3.3)	3.2 (1.7–6.2)
Leaks ≥2 times/day vs. leaks <2 times/day or no leaking	13.8 (14.0)	2.3 (2.2)	7.4 (3.6–15.2)
Wears pads to stay dry¶	28.1 (28.3)	2.6 (2.5)	15.5 (7.7–31.0)
Frequent urination >½ time vs. frequent urination ≤½ time	11.1 (10.9)	10.4 (10.8)	1.0 (0.6–1.7)
Bothered by dripping or leaking urine¶	11.2 (11.7)	2.3 (2.0)	6.6 (2.8–15.4)
<b>Bowel#</b>			
Diarrhea	20.9 (22.1)	37.2 (33.2)	0.50 (0.34–0.72)
Painful bowel movements	9.2 (10.7)	13.6 (10.6)	1.0 (0.58–1.8)
Bowel urgency	14.5 (16.1)	35.7 (30.5)	0.40 (0.27–0.59)
Wetness in rectal area	14.2 (14.7)	21.8 (20.7)	0.63 (0.40–0.99)
Painful hemorrhoids	10.3 (9.5)	16.3 (19.3)	0.38 (0.23–0.64)
Bothered by frequent bowel movement, pain, or urgency¶	3.3 (4.1)	8.4 (5.7)	0.68 (0.31–1.5)
<b>Sexual</b>			
No/little vs. some/a lot of interest in sexual activity	42.8 (45.8)	51.0 (43.2)	1.1 (0.79–1.6)
No sexual activity vs. any sexual activity	46.3 (49.9)	45.5 (35.4)	2.4 (1.6–3.5)
Erection insufficient for intercourse	79.6 (82.1)	61.5 (50.3)	6.4 (4.2–9.6)
Bothered by sexual dysfunction  ,**			
Age 55–59 y	59.4 (74.9)	25.3 (39.9)	5.0 (1.7–14.7)
Age 60–74 y	53.2 (52.8)	46.1 (46.6)	1.3 (0.9–1.9)

Table 3 in Potosky et al. (2000) p. 1586

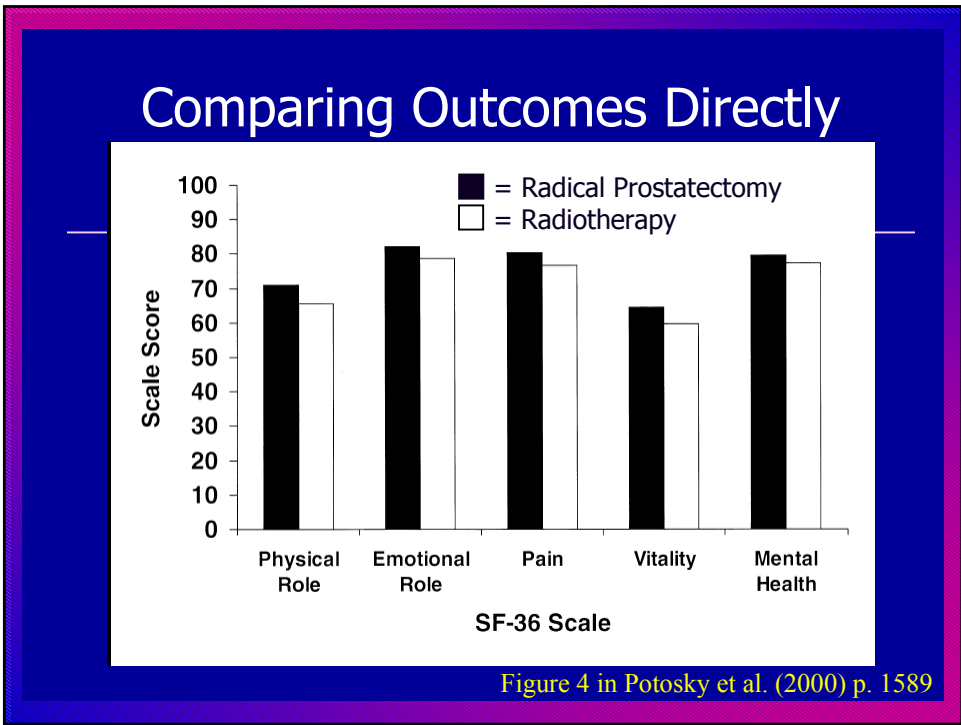


Figure 4 in Potosky et al. (2000) p. 1589

Discussion Topics

## Matching vs. Multivariate Adjustment

- Whether or not matched sampling is used, further analytical adjustments may be desirable to control residual bias and to increase efficiency.
- Rubin (1979, JASA) found that regression adjustment of matched-pair differences is a robust technique.

## What should always be done ... and often isn't?

- Collect data so as to be able to model selection
- Demonstrate selection bias – need for PS
- Ensure covariate overlap / specify relevant population carefully in light of overlap
- Evaluate covariate balance after PS application
- Model or estimate treatment effect in light of PS adjustment / matching
- Estimate sensitivity of results to potential hidden bias or uncaptured selection bias