For *Objective* Causal Inference, Design Trumps Analysis

Donald B. Rubin

Harvard University

Statistics 140, W.G. Cochran, 1968
Classical Experimental Design

Clear separation between:

- "Science" (and …)

		X	Y(0)	Y(1)
Units	1			
	N			

X = Covariates unaffected by treatments

Y(0) = Potential outcomes under control treatment

Y(1) = Potential outcomes under active treatment

Notation due to Neyman (1923) in context of randomized experiments

Classical Experimental Design

- Clear separation between Science and what we do to learn about Science:
  - Randomized assignment of treatments
  - W = Vector of N treatment indicators
- This distinction and the consequential clarity should be maintained and should not be forgotten when designing observational studies
- Cochran & Stat 140, great advantage to start here

Causal Inference is a Missing Data Problem

RCM (Holland, 1986) for work in 1970's

Maintains critical distinction from experimental design

- Same notation for science whether try to learn about it from randomized experiment or observational study.
- Earlier, "Y_{obs}" used in nonrandomized studies with W a predictor in regressions, paths, arrows

  $Y_{obs} = \{Y_{obs,i}\}$

  $Y_{obs,i} = W_i Y_i(1) + (1-W_i)Y_i(0)$

  Entangles Science and assignments

Assignment Mechanism

Creates missing potential outcomes

$$\Pr(W|X,Y(0),Y(1)) \qquad \text{(AM)}$$

Randomized experiments are:

unconfounded: AM = $\Pr(W|X)$, and

probabilistic: $1 > \Pr(W_i|X_i) > 0$ for all i

Earlier, words describing AM but no explicit mathematical notation or expressions (e.g., Roy, 1953)

Same is true for potential outcomes before Neyman's (1923) notation (e.g., Fisher, 1918)

Design Observational Studies to Approximate Randomized Trials

1. Hide outcome data until the design phase is complete
2. Think very carefully about decision makers and the key covariates that were used to make treatment decisions
3. If key covariates are not observed or very noisy, usually best to give up and seek better data source
4. Find subgroups (subclasses or matched pairs) in which the treatment and control groups have balance – essentially the same distribution of observed covariates
   - Not always possible to achieve balance
   - Inferences are limited to subgroups where balance is achieved

- #1 - #4 combine to create an objective design that approximates a randomized trial in each subclass that is balanced with respect to observed covariates

Cochran (1968) – Illustrative Example with One Key Covariate

- Population: Male smokers in U.S.
- Treatment = cigar/pipe smoking
- Control = cigarette smoking
- Outcome = death rate/1000 person years
- Decision maker is the individual male smoker
- Reason for a smoking male to choose cigarettes versus cigar/pipe?
- **Age** is a key covariate for selection of smoking type for males

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

Comparison of Mortality Rates for Two Smoking Groups in U.S.

	Cigarette Smokers	Cigar/Pipe Smokers
Mortality Rates per 1000 person-years, %	13.5	17.4
Adjusted Mortality Rates using subclasses, %		
2 age subclasses	16.4	14.9
3 age subclasses	17.7	14.2
9-11 age subclasses	21.2	13.7

Source: Cochran WG. The effectiveness of adjustment of subclassification in removing bias in observational studies. Biometrics 1968; 24:295-313.

But 20 four-class covariates $\Rightarrow$ over million million subclasses

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

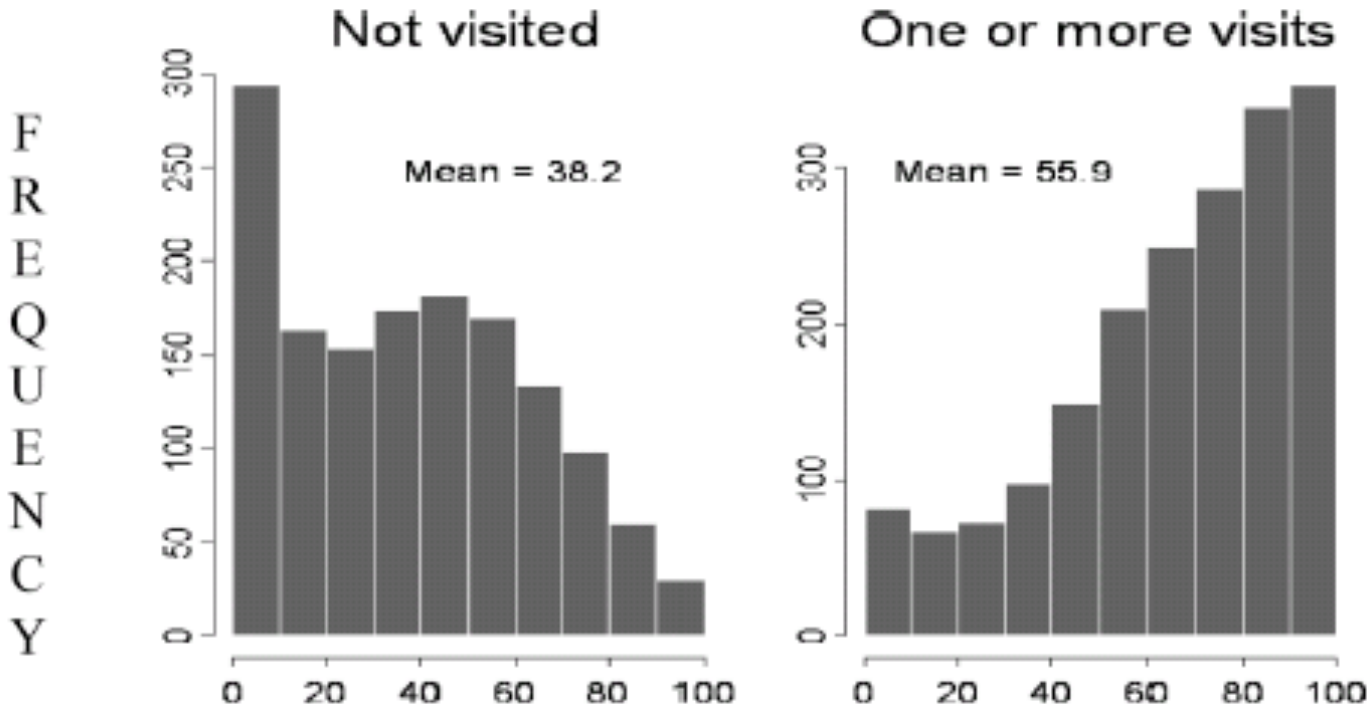
Propensity Score Subclass	Women		Estimated Survival Rate for Women		Estimated Causal Effect
	Breast Conservation (BC)	Mastectomy (Mas)	BC	Mas	BC – Mas
	n	n	%	%	%
1	56	1008	85.6	86.7	-1.1
2	106	964	82.8	83.4	-0.6
3	193	866	85.2	88.8	-3.6
4	289	978	88.7	87.3	1.4
5	462	604	89.0	88.5	0.5
Averages Across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

Diagnostics for Accessing Balance

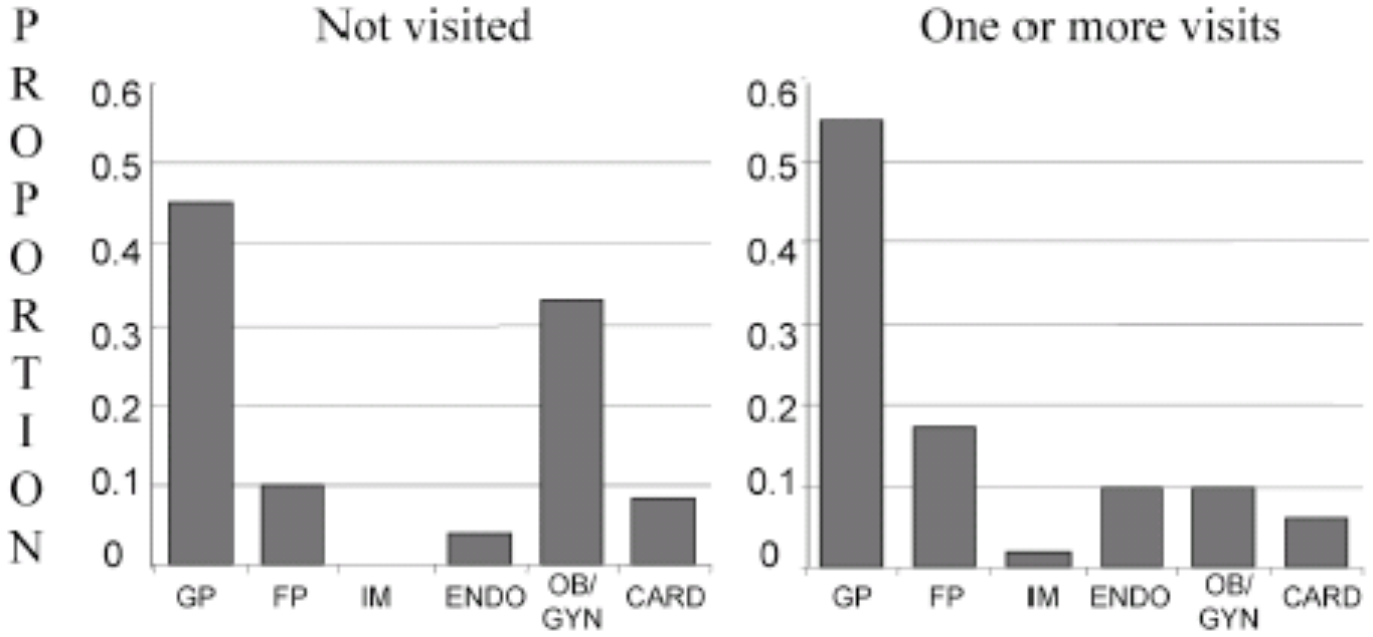
- Assessing balance simpler in large samples, just as with randomized experiments
- To illustrate diagnostics, use a marketing application that involved a weight loss drug
- Units = doctors
- Treatment = sales rep “visits” doctor to discuss
- Control = no visit
- Decision-makers = sales reps
- Key covariates = prior Rxs, medical specialty, years in practice, size of practice, etc.

Histograms for background variable: Prior Rx Score (0-100) at Baseline



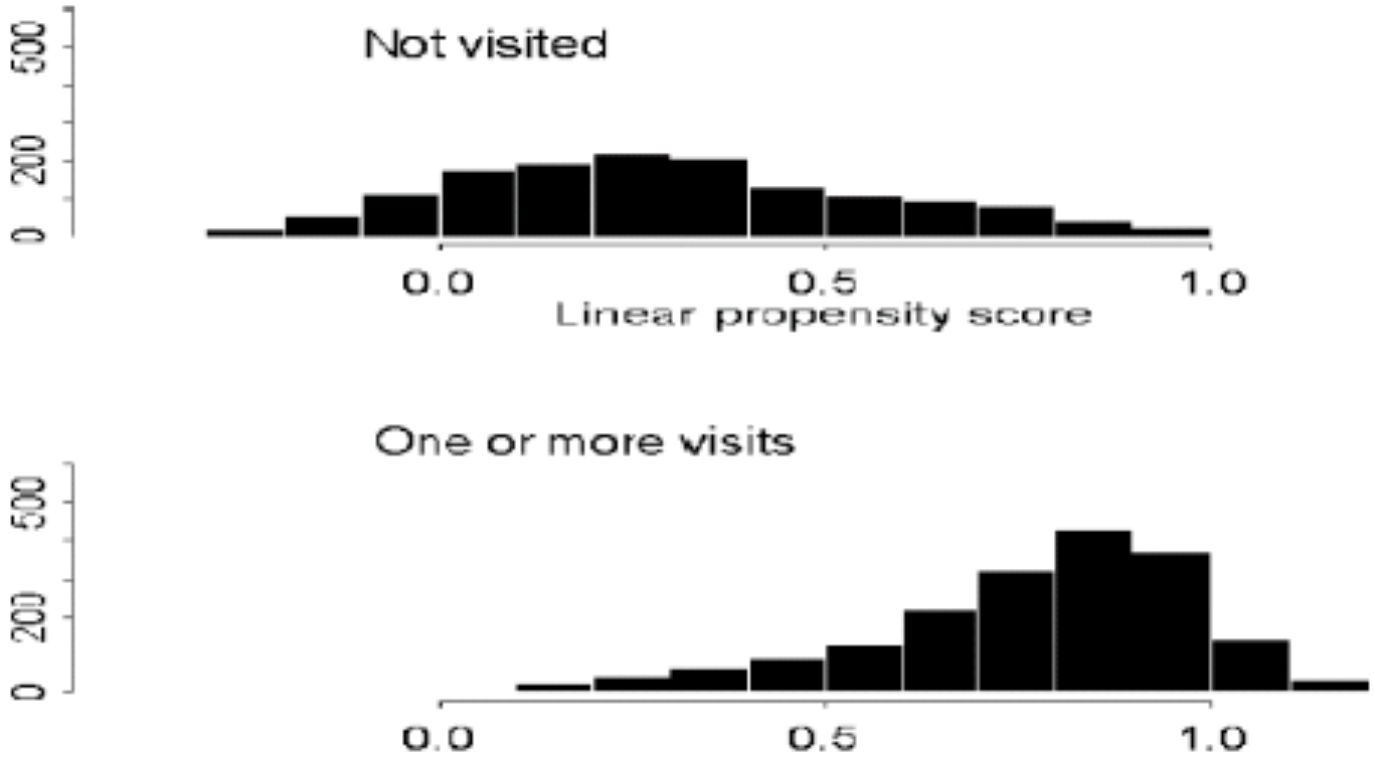
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for background variable: Specialty



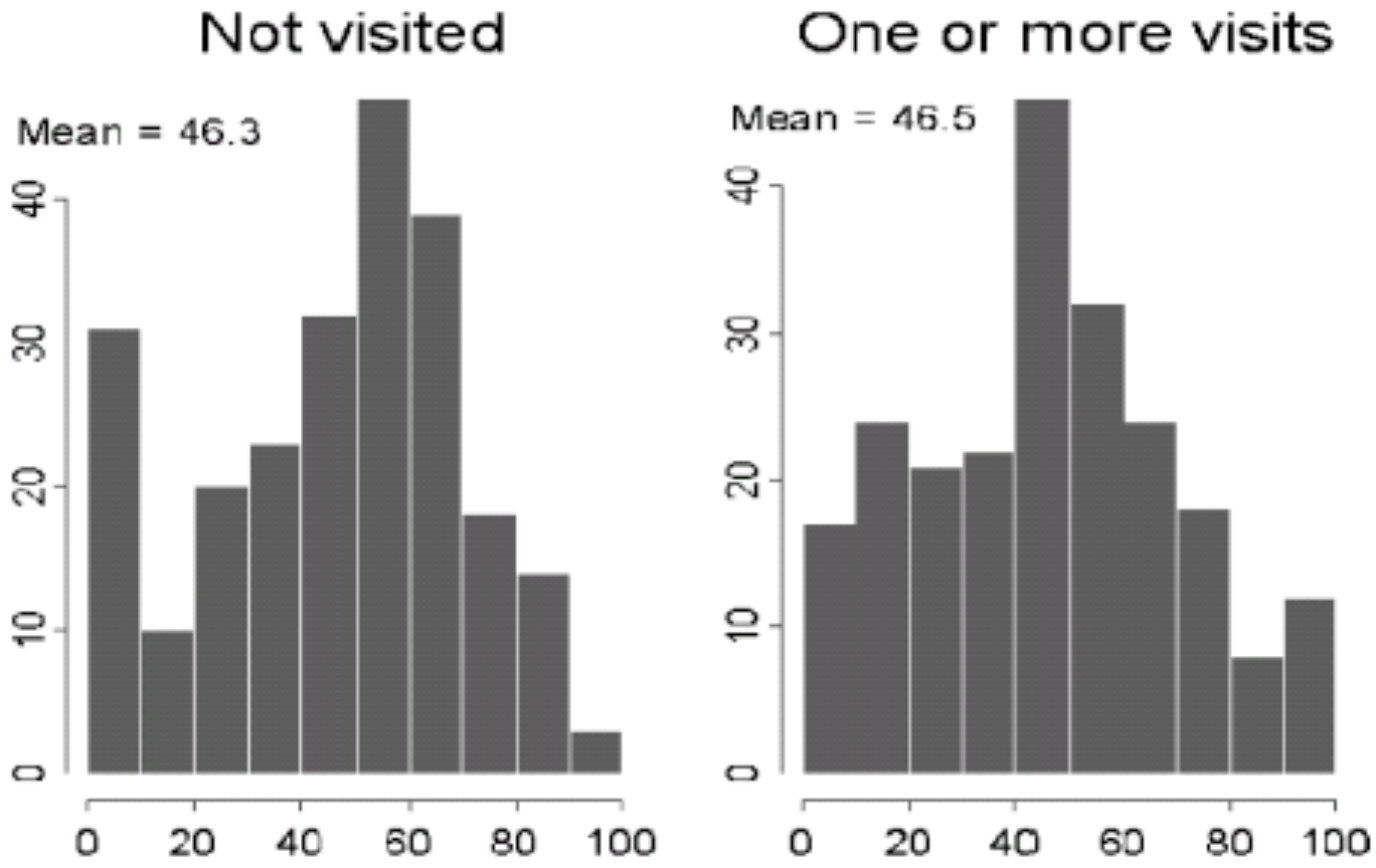
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for summarized background variables: Linear Propensity Score



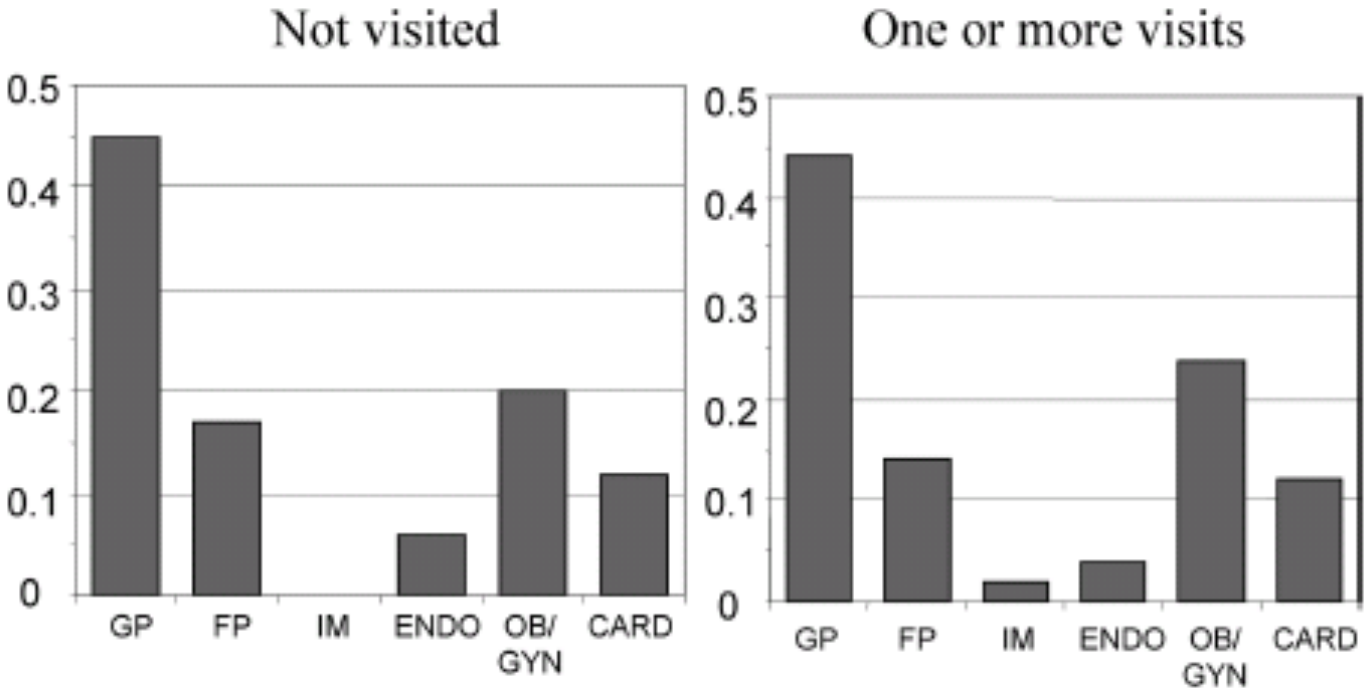
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for a variable in a subclass of propensity scores: Prior Rx Score



Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for a variable in a subclass of propensity scores: Specialty



Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Marketing Example: Achieved Balance

- Within each narrow subclass of propensity scores, the treatment and control groups will be as balanced as if randomly divided
- Claim: This holds for all subclasses in which there are both treated and control subjects, and holds for all covariates that were used to estimate the propensity score
- Works best when the propensity score subclasses have large sample sizes and are relatively narrow
- Five to ten propensity score subclasses often fully adequate to balance all covariates
- No outcome data used in the design stage

Simple Noncompliance, Instrumental Variables, and Bayesian Generalizations

- Template for other observational studies involves more complex randomized experiment
- Illustrate with completely randomized experiment with noncompliance with assigned treatment
- Return later to combined analysis with observational study design

Sommer and Zeger Vitamin A Data

Row	True Compliance Type	Treatment Assignment	Treatment Received	Y_{obs}	Number of Children
1	?	0	0	0	11514
2	?	0	0	1	74
3	N	1	0	0	2385
4	N	1	0	1	34
5	C	1	1	0	9663
6	C	1	1	1	12
					23682

Reference: Sommer and Zeger (1991). On Estimating Efficacy from Clinical Trials. Statistics in Medicine.

Results of Three Standard MoM Analyses

Method	Estimate	Calculation	Row Comparison
ITT	-0.0026	$= \frac{12+34}{9663+2385+12+34} - \frac{74}{11514+74}$	3, 4, 5, & 6 vs. 1 & 2
As-treated	-0.0065	$= \frac{12}{9663+12} - \frac{34+74}{11514+2385+34+74}$	5 & 6 vs. 1, 2, 3, &4
Per protocol	-0.0052	$= \frac{12}{9663+12} - \frac{74}{11514+74}$	5 & 6 vs. 1 & 2

Reference: Sommer and Zeger (1991). On Estimating Efficacy from Clinical Trials. Statistics in Medicine.

MoM CACE Analysis

$ACE = p_N \cdot NACE + p_C \cdot CACE$

$-0.0025 = 0.2 \cdot NACE + 0.8 \cdot CACE$

$-0.0025 = 0.8 \cdot CACE \rightarrow CACE = -0.0025/0.8 = -0.0031$

Bayesian Analysis of Sommer & Zeger Data

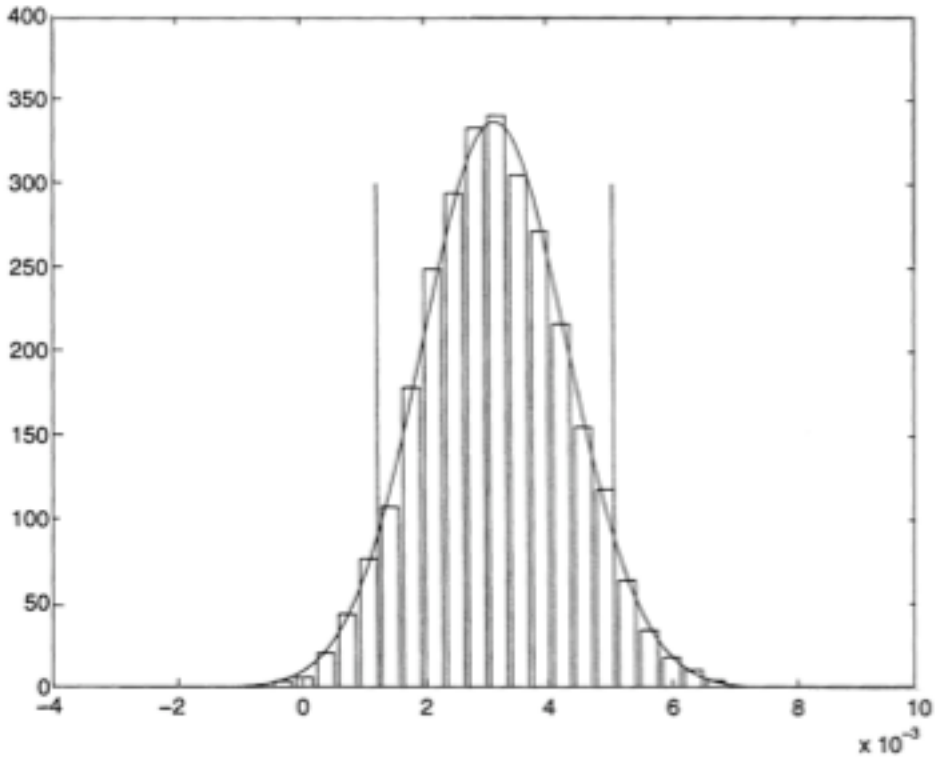


FIG. 3. *Histogram of CACE with exclusion restriction (data from Table 3).*

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Bayesian Analysis of Sommer & Zeger Data, Marginal Posterior Distributions with and without Exclusion Restriction

Estimand	Exclusion restriction	Mean	Standard deviation	Median	5^{th} percentile	95^{th} percentile
CACE	No	3.1	2.5	3.2	-0.9	7.0
$ITT_Y^{(n)}$	No	0.5	10.1	0.2	-14.1	17.5
CACE	Yes	3.1	1.2	3.1	1.2	5.1

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Bayesian Analysis of Sommer & Zeger Data

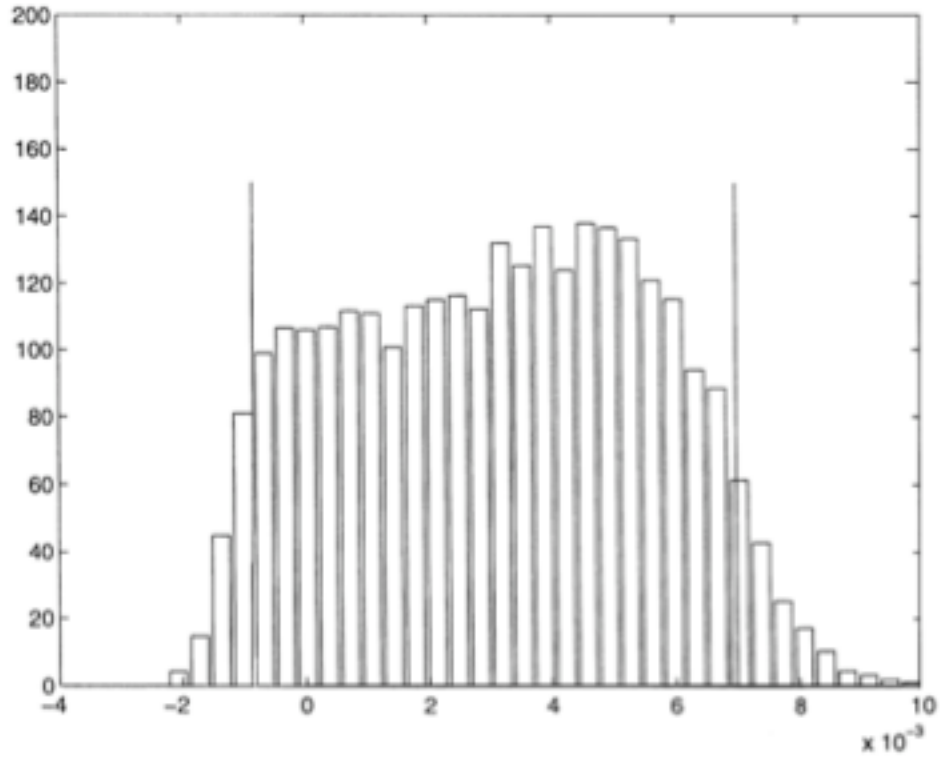


FIG. 1. *Histogram of CACE without exclusion restriction (data from Table 3).*

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Bayesian Analysis of Sommer & Zeger Data

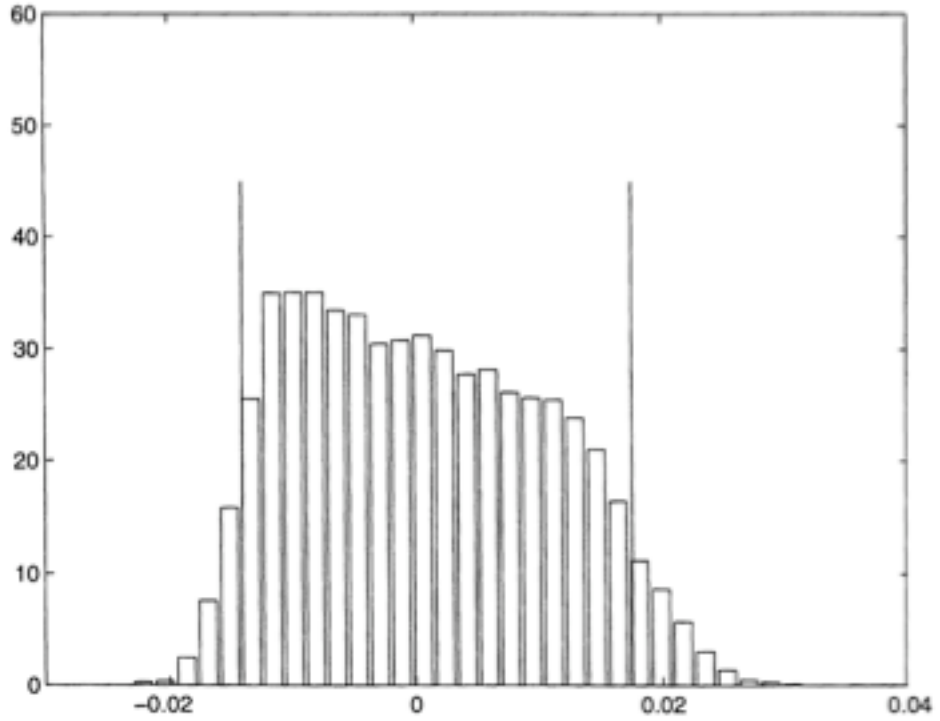


FIG. 2. *Histogram of* $\mathrm{ITT}_Y^{(n)}$ *without exclusion restriction (data from Table 3).*

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Bayesian Analysis of Sommer & Zeger Data

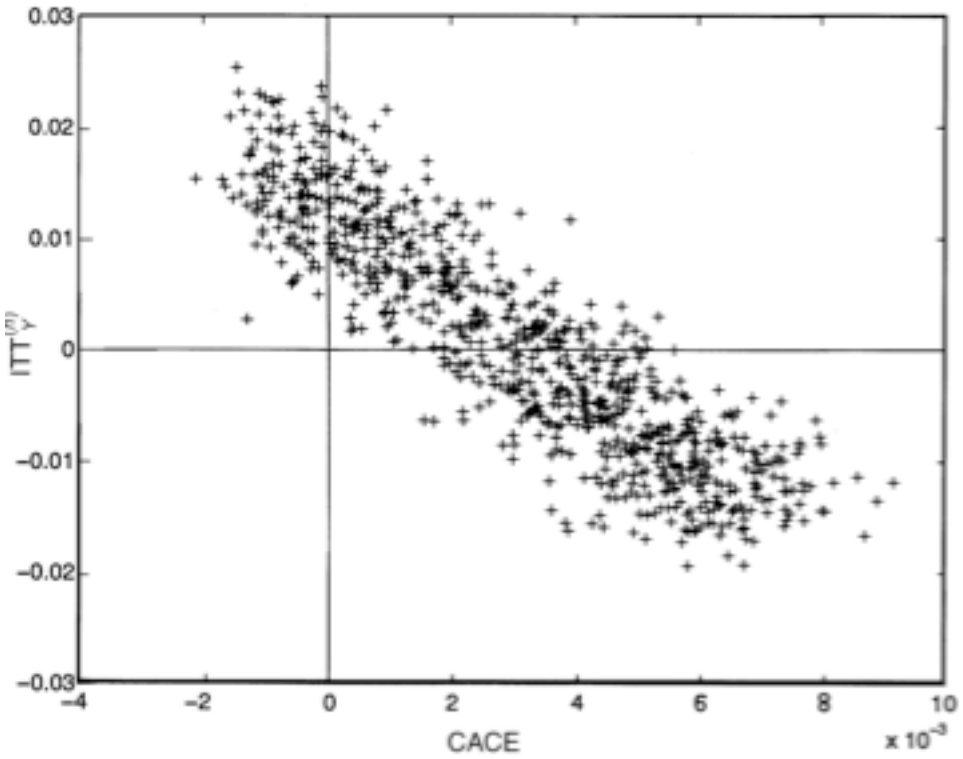


FIG. 4. *Joint posterior distribution of CACE and* $\mathrm{ITT}_Y^{(s)}$ *(data from Table 3).*

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, Population Parameters with Exclusion Restrictions and Monotonicity

T	$P(C_i = t \mid \pi)$	$D_i(0)$	$D_i(1)$	$Y_i \mid C_i = t, Z_i = 0, \pi$	$Y_i \mid C_i = t, Z_i = 0, \pi$
c	0.25	0	1	$N(0.1, 0.16)$	$N(0.9, 0.49)$
n	0.45	0	0	$N(1.0, 0.25)$	$N(1.0, 0.25)$
a	0.30	1	1	$N(0.0, 0.36)$	$N(0.0, 0.36)$

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, One Sample

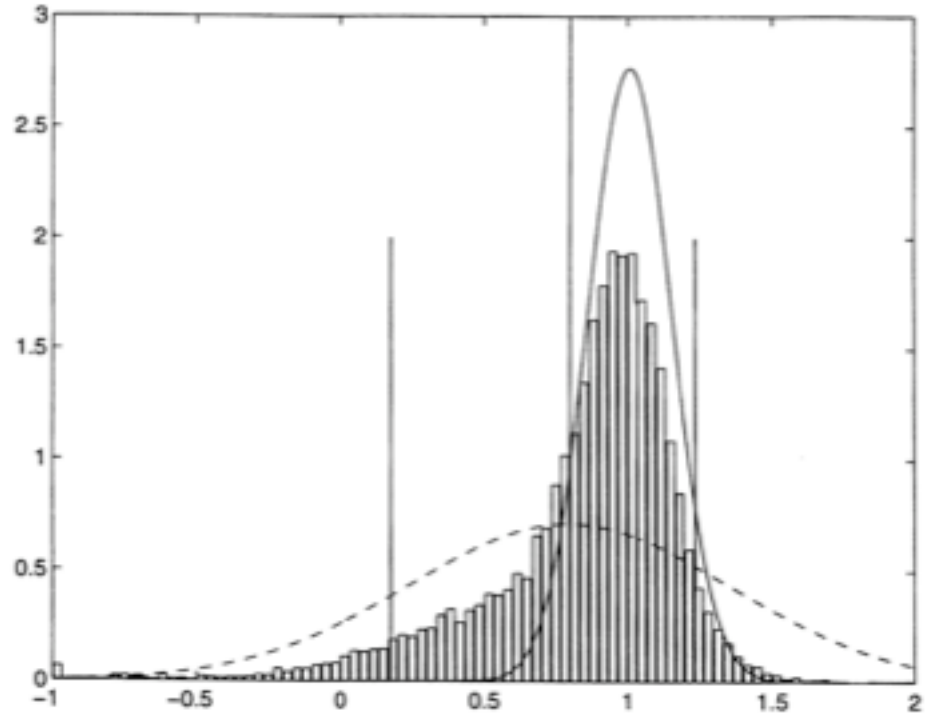


FIG. 5. *Estimates of the posterior distribution of CACE under exclusion restriction and monotonicity condition (data analyzed in Table 6): histogram is based on simulation, solid line is normal approximation based on mle, dashed line is normal approximation based on* $\widehat{\text{IVE}}$.

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

Hypothetical Example Illustrating Frequentist Superiority of Bayes over IVE (MoM) and MLE, Frequentist Evaluation under Monotonicity and Exclusion Restrictions

Estimator	Mean bias	Median bias	Root mean squared error	Median absolute error	90% interval	
					Coverage rate	Median width
Posterior mean	-0.10	-0.07	0.48	0.30	0.91	1.61
Posterior median	-0.08	-0.06	0.51	0.32		
MLE	-0.14	-0.12	0.51	0.31	0.74	1.11
IVE	0.55	0.13	2.31	0.54	0.91	2.78

Imbens G.W. and Rubin D.B. (1997) Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance. Annals of Statistics 25(1):305-327.

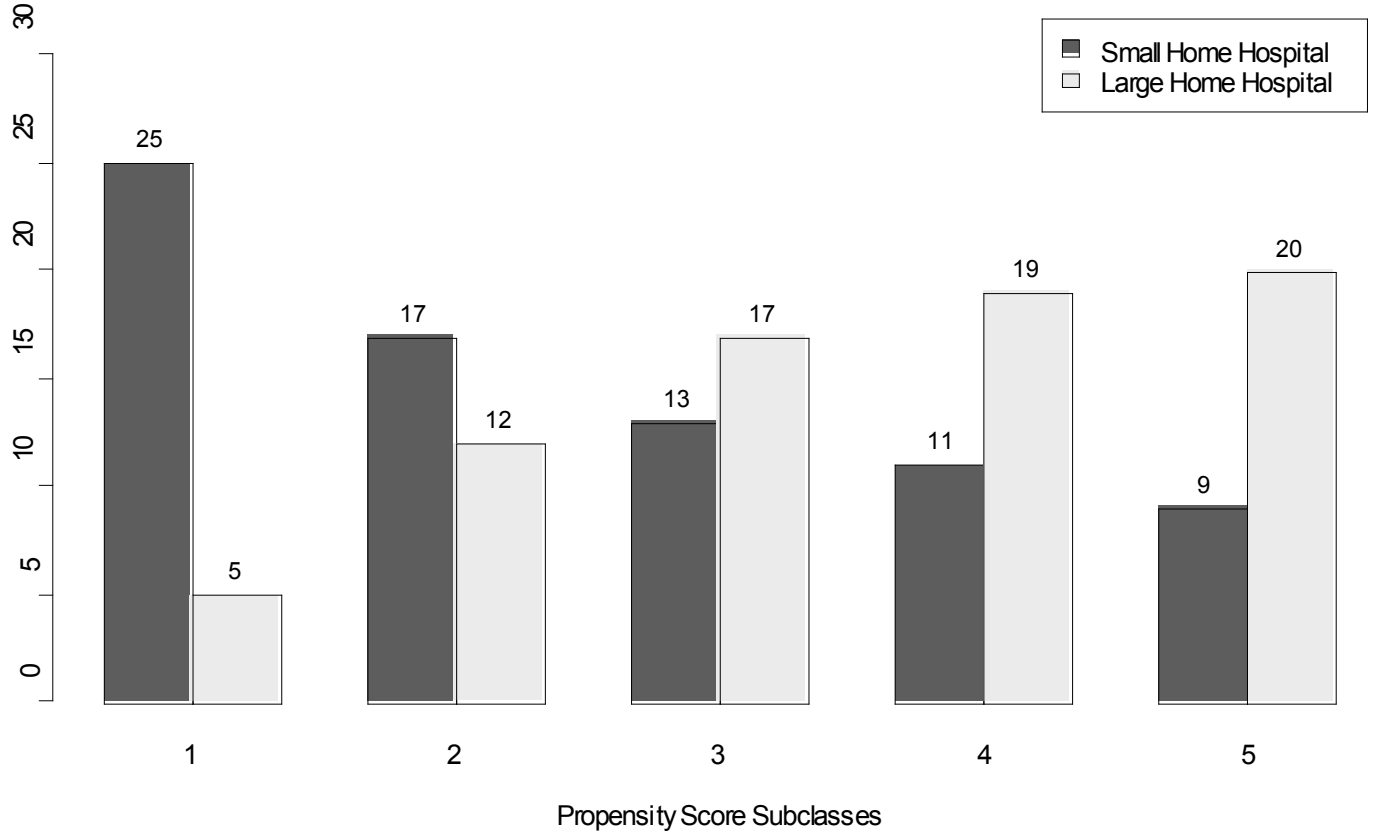
Using a More Complex Template: Randomized Block Experiment with Noncompliance

- Causal effect of Large versus Small treating hospitals on cardia cancer survival
- Dataset from Karolinska Institute, Stockholm
- Medical researchers accept unconfounded assignment of "home (diagnosing) hospital" type, but NOT treating hospital type because of self-selected transfers
- Consider transfers between hospital types as a form of noncompliance with assignment

Using a More Complex Template: Randomized Block Experiment with Noncompliance

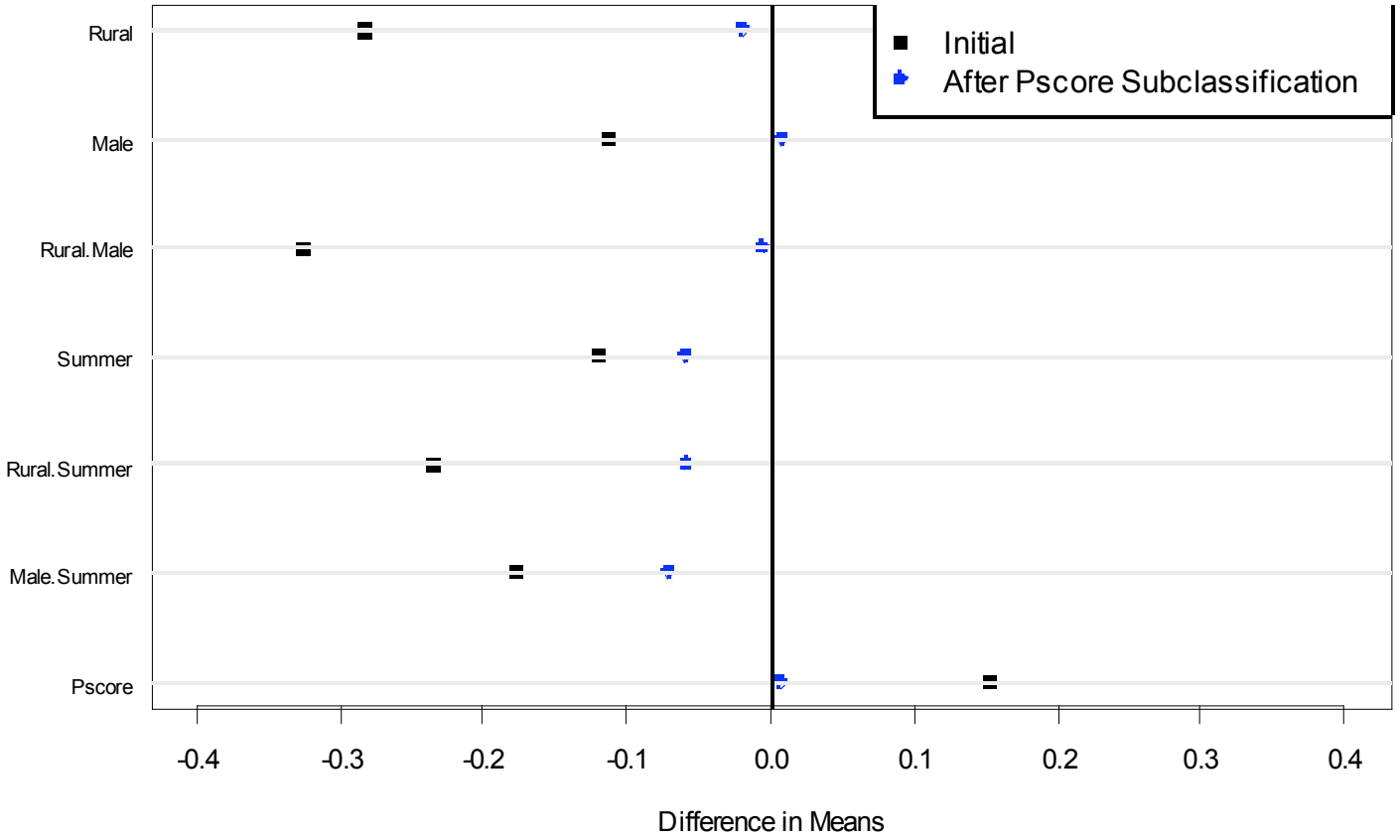
- Design has two distinct phases
- Phase 1, no outcome data available:
  - Propensity score analysis to approximate randomized block experiment for home hospital type
  - Ensure subclassification can create balance on covariates for large and small home hospital types
- Phase 2, uses intermediate outcome data on transfers:
  - Outline of the analysis for estimating causal effect of treating hospital type
  - Ensure within each subclass that there appear to be compliers who are treated in both and large and small treating hospitals

Figure 5.1: Cardia Cancer, Number of People, Subclassified by Propensity Score



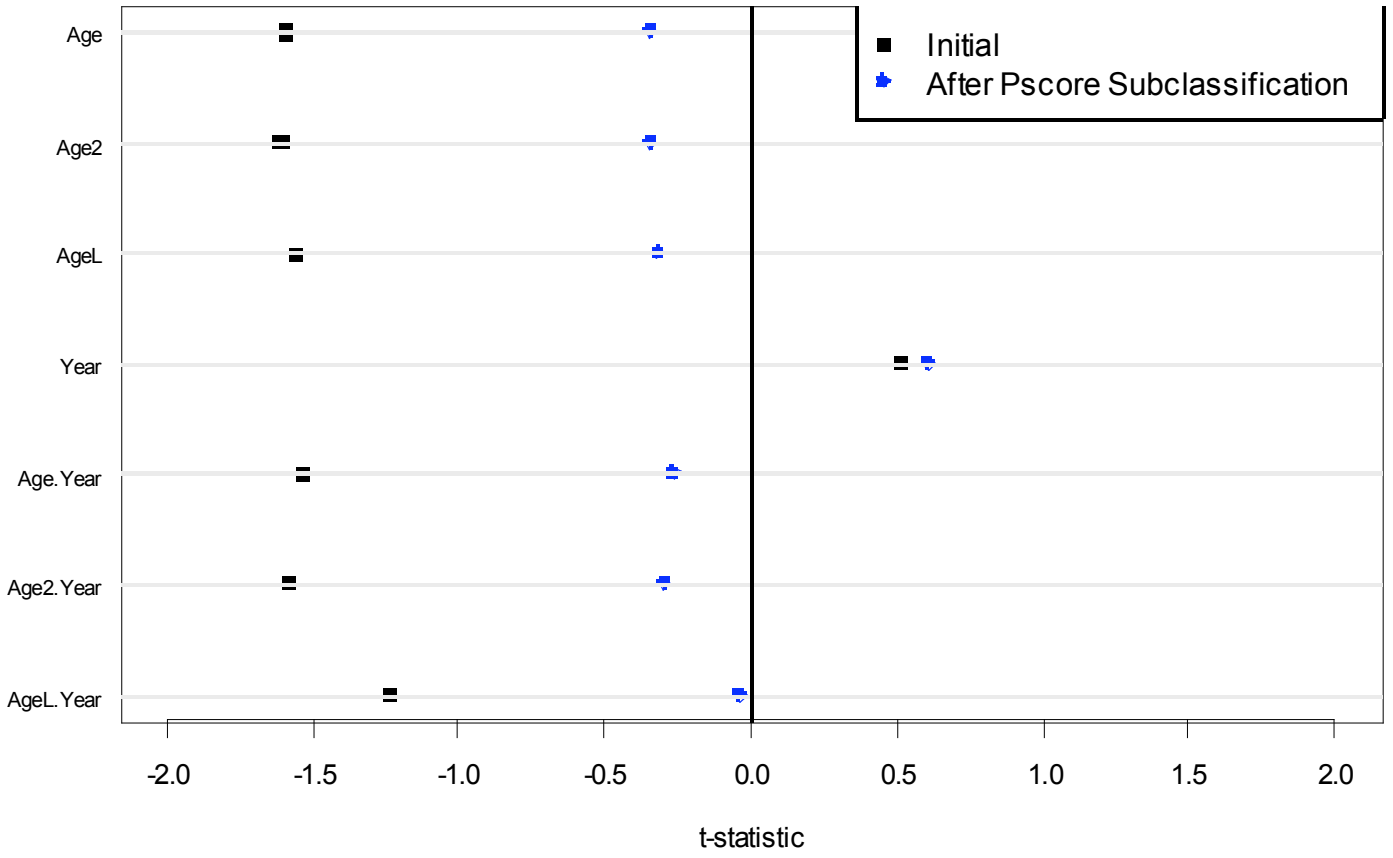
Reference Rubin, D.B. For Objective Causal Inference, Design Trumps Analysis. Annals of Applied Statistics, 2008.

Figure 5.2: Cardia Cancer, Difference in Means for Binary Covariates and Pscore



Reference Rubin, D.B. For Objective Causal Inference, Design Trumps Analysis. Annals of Applied Statistics, 2008.

Figure 5.3: Cardia Cancer, t-statistics for Continuous Covariates



Reference Rubin, D.B. For Objective Causal Inference, Design Trumps Analysis. Annals of Applied Statistics, 2008.

Outline of Analysis within Each Subclass

- Critical that we anticipate compliers in both large and small treating hospitals within each subclass
- Monotonicity assumption = no defiers; medically very plausible

$$\text{ITT} = \pi_{LS}\,\text{ITT}_{LS} + \pi_{SS}\,\text{ITT}_{SS} + \pi_{LL}\,\text{ITT}_{LL}$$

$$CACE \equiv ITT_{LS} = \frac{1}{N_{LS}} \sum_{i \in LS} (Y_i(L) - Y_i(S)),$$

$$ITT = \pi_{LS} ITT_{LS}, \; ITT_{LS} = ITT / \pi_{LS}.$$

Method of Moments Estimates of the Number of Compliers Treated in Large and Small Hospital Types Under Monotonicity

"Assigned"/Randomized Home Hospital Type	**Treating Hospital Type**	**Approximate N in LS Principal Stratum Subclass**				
h	**T**	**1**	**2**	**3**	**4**	**5**
ℓ	L	3	3	11	9	5
s	S	14	5	8	5	2

Table 5.3: Cardia Cancer: Observed Counts in Observed Groups and Approximate Counts in Principal Strata Under Monotonicity Assumption – Subclass 3

(1)			(2)	(3)	(4)	(5)	(6)
"Assigned"/Randomized Home Hospital Type			Treating Hospital Type *T*	#	Underlying Principal Strata: *h* =	Approximate Proportion in Population in Principal Strata	Approximate N in *LS* Principal Stratum
	h	#			ℓ s		
(1) (2)	ℓ	17	L	17	L L L S	38% 62%	11
			S	0	S S	0%	
(3) (4)	*s*	13	L	5	L L	38%	
			S	8	S S L S	0% 62%	8

Reference Rubin, D.B. For Objective Causal Inference, Design Trumps Analysis. Annals of Applied Statistics, 2008.

Summary: Objective Observational Study Design

- Should approximate a randomized experiment
  - No ultimate outcome data used or examined – "prospective"
  - Carefully consider decision-makers and the covariates used to make treatment assignments
  - If dataset is missing key covariates, usually do NOT continue
  - Use propensity score estimation to help create subclasses or matched pairs that achieve "balance" on covariates
  - Balance means treated and control subjects have distributions of covariates that are at least as similar as if they had been randomized into treatment and control
  - Analysis takes place within each subclass, and then answers are combined across subclasses