General Considerations When Designing Observational Studies

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
5. Protocol specified analysis

- #1 - #5 combine to create an objective design that approximates a randomized trial in each subclass that is balanced with respect to observed covariates

Illustrative Example with One Key Covariate (Cochran, 1968)

- 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

Reference: Rubin DB. The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels With The Design of Randomized Trials. Statistics in Medicine 2007

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

Variable	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.

Note: 20 four-level covariates ⇒ 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 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 in estimating the propensity score
- Using diagnostics to assess and to document balance is critical

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 general practice?
- Observational data available
  - SEER Database: covariates, treatments, post-surgery outcomes
- Design phase
  - Hide outcomes
  - Balance covariates between treatment and control
- Reasons for mastectomy versus breast conservation
  - Age, marital status, region of country, urbanization, race, size of tumor, etc.

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 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.

Propensity Score Analysis Approach

- Estimate propensity scores
- Then subclassify (or match) on propensity score as if the only covariate, e.g., 5-10 subclasses
- Why does this work?
  - Creates balance in each subclass on **ALL** covariates used in estimating the propensity score
  - This balance will be achieved in <u>large</u> samples just like the balance that will be achieved in a large randomized clinical trial

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

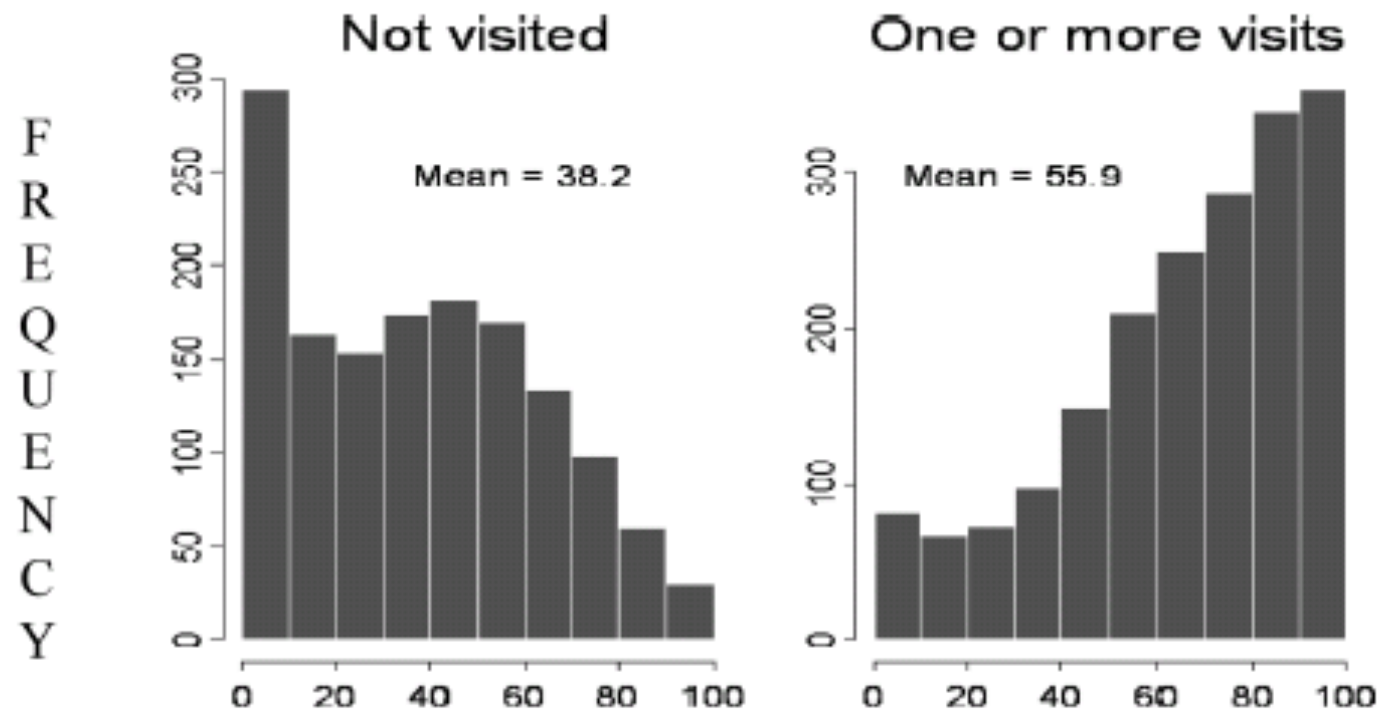
Propensity Score Subclass	Women: Breast Conservation (BC)	Women: Mastectomy (Mas)	Estimated Survival Rate for Women: BC	Estimated Survival Rate for Women: Mas	Estimated Causal Effect: BC – Mas
	n	n	%	%	%
1	56	1008	85.6	86.7	-1.1
2	106	964	82.8	83.4	-0.6
3	193	866	85.2	88.8	-3.6
4	289	978	88.7	87.3	1.4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

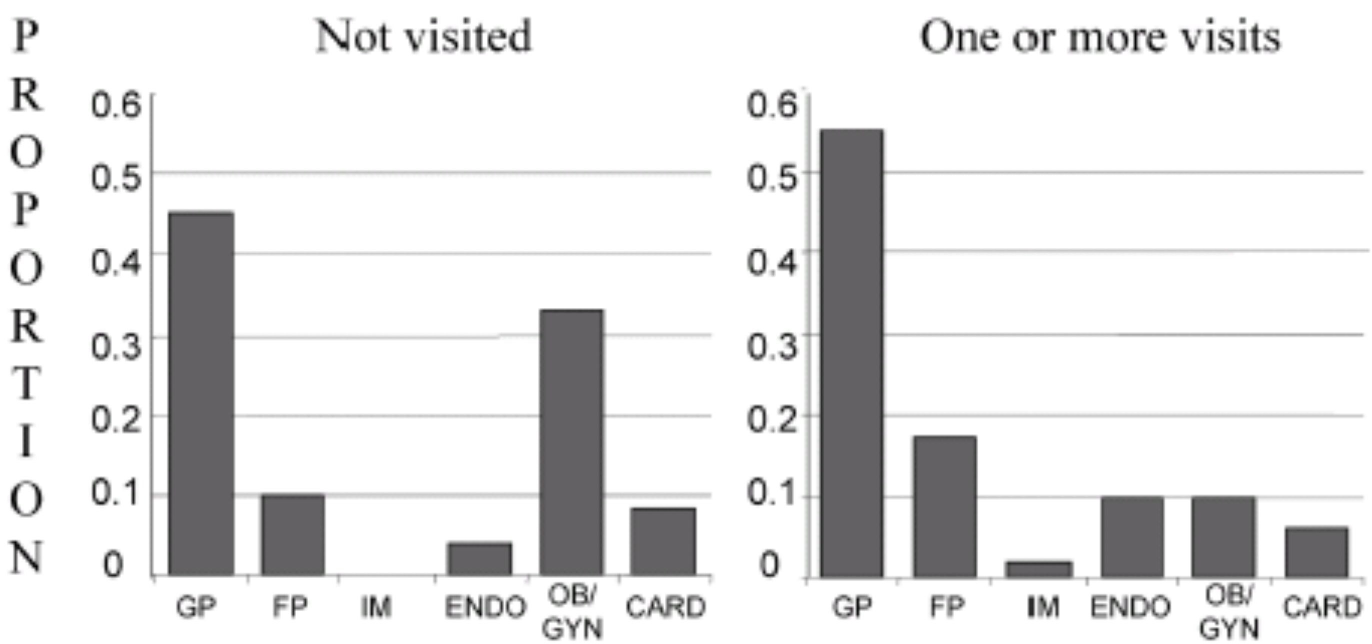
- Assessing balance simpler in large samples, just as with randomized experiments
- To illustrate diagnostics, use a marketing application that involved a weight loss drug
- Units = doctors
- Treatment = sales rep "visits" doctor to discuss
- Control = no visit
- Decision-makers = sales reps
- Key covariates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

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



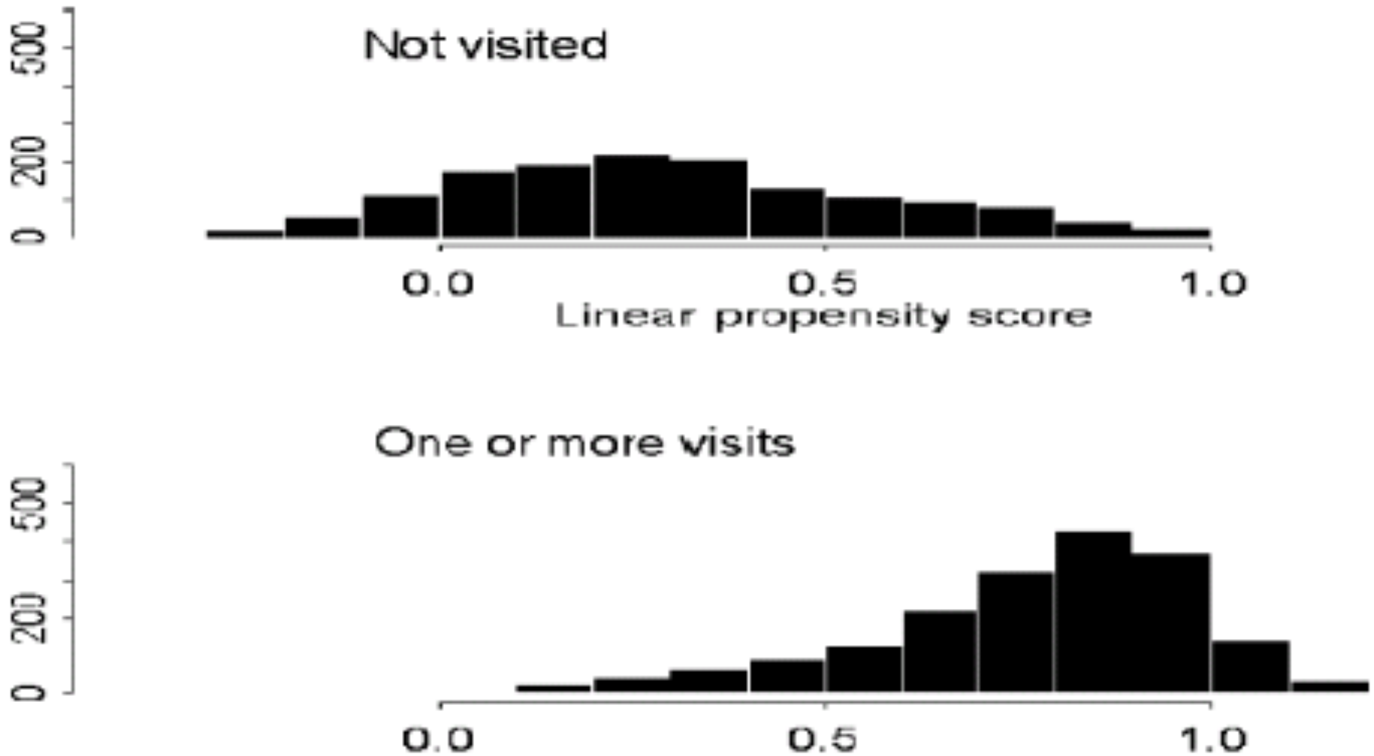
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for background variable: Specialty



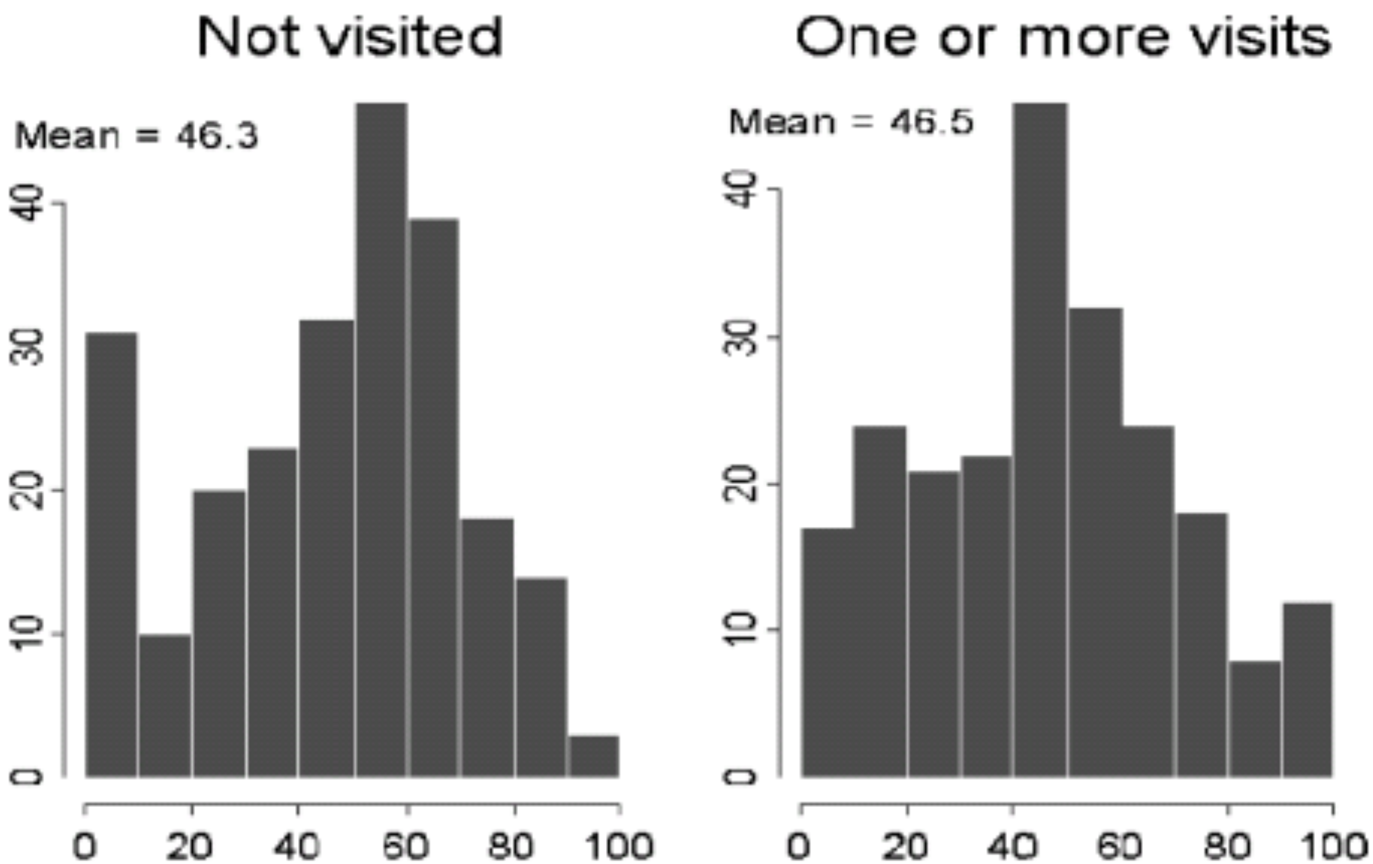
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

Histograms for summarized background variables: Linear Propensity Score



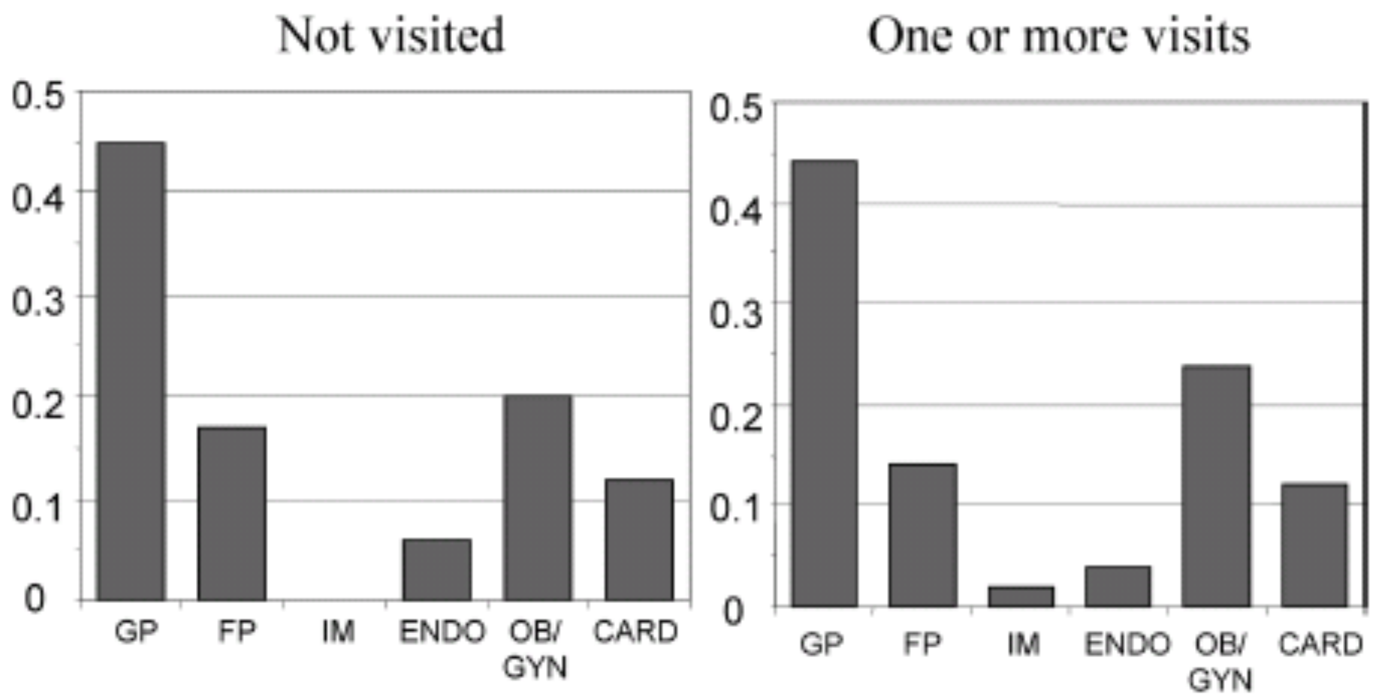
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

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.

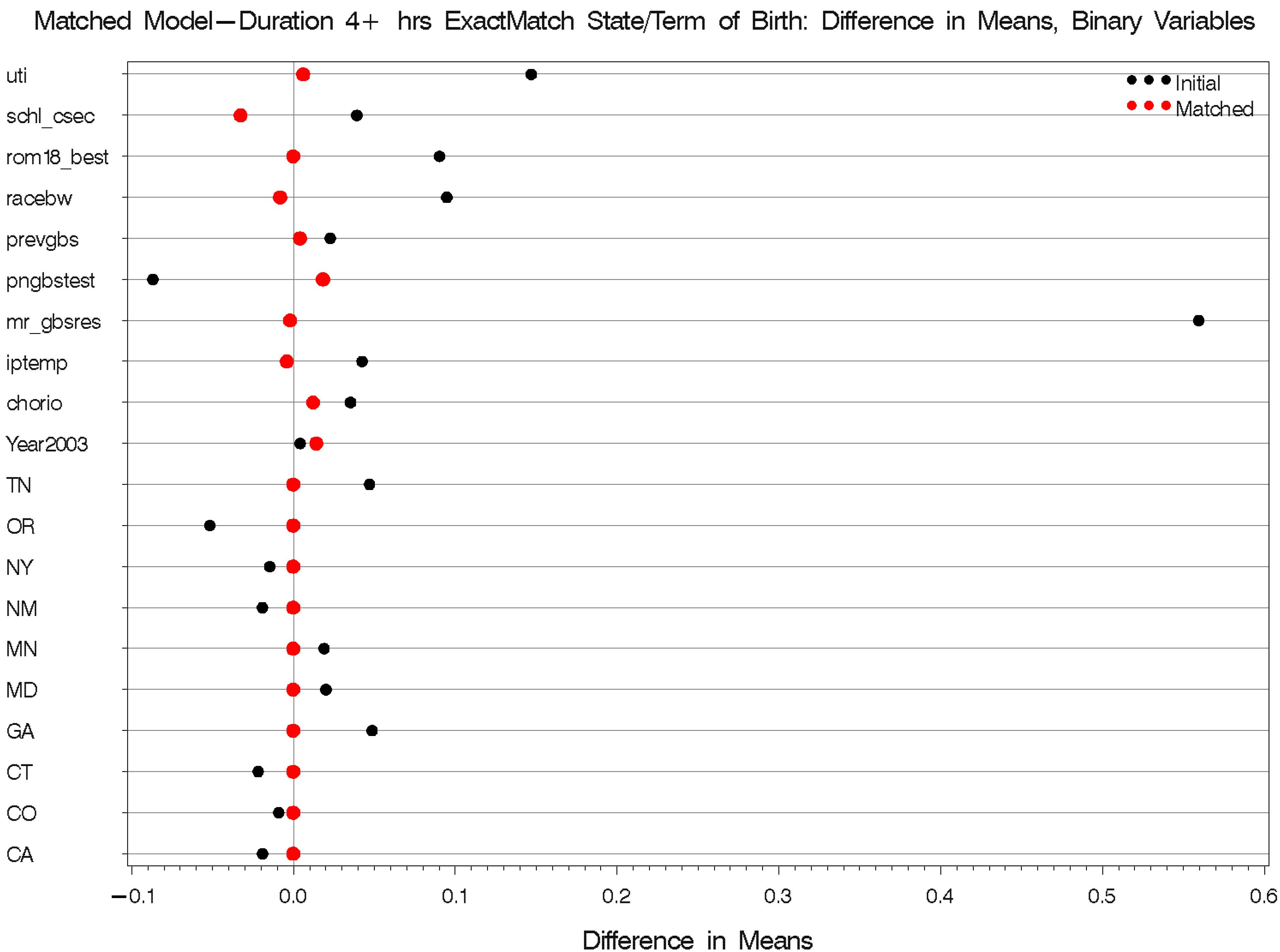
Histograms for a variable in a subclass of propensity scores: Specialty



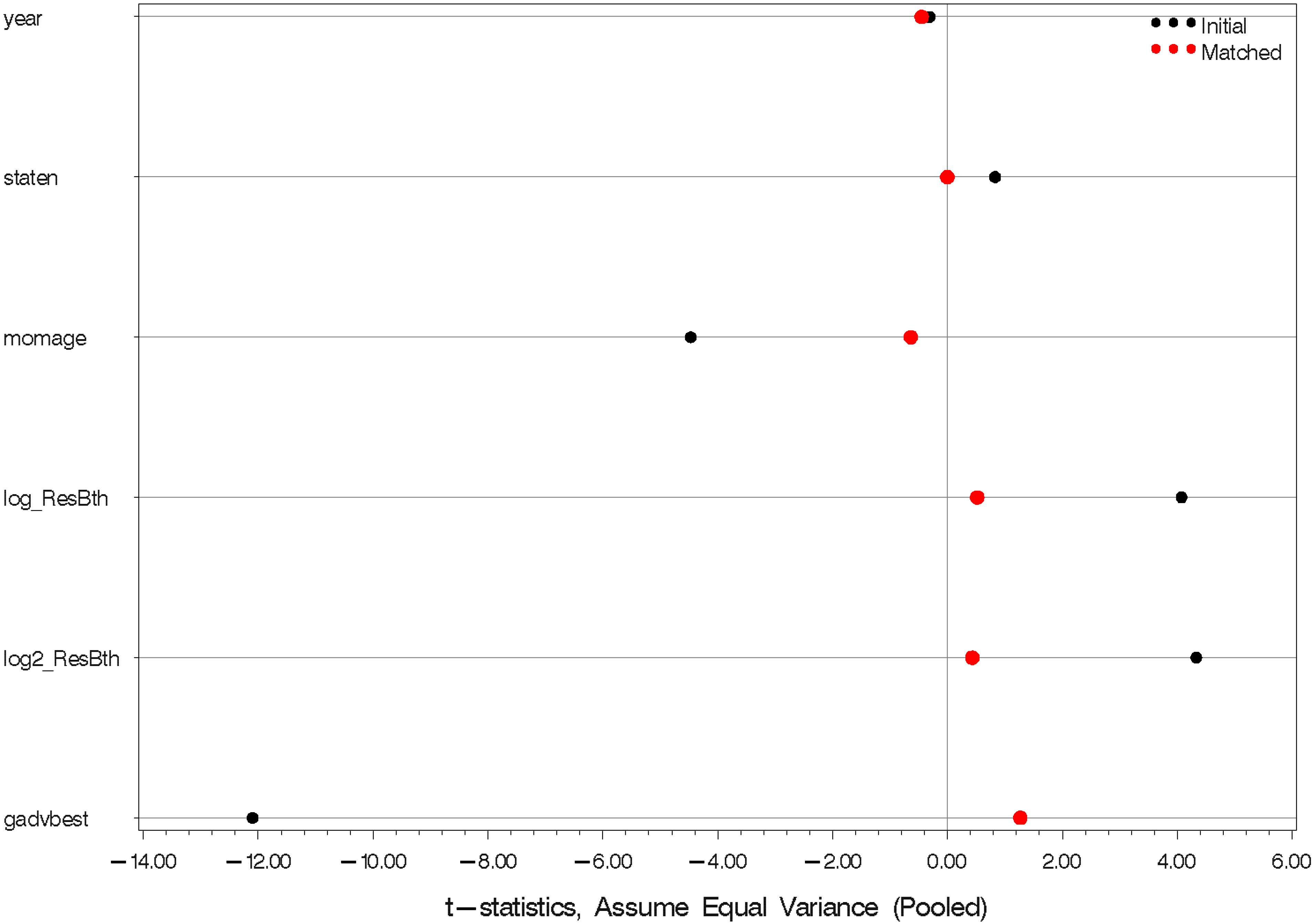
Source: Rubin DB and Waterman RP. Estimating Causal Effects of Marketing Interventions Using Propensity Score Methodology. Statistical Science 2006; 21(2):206-222.

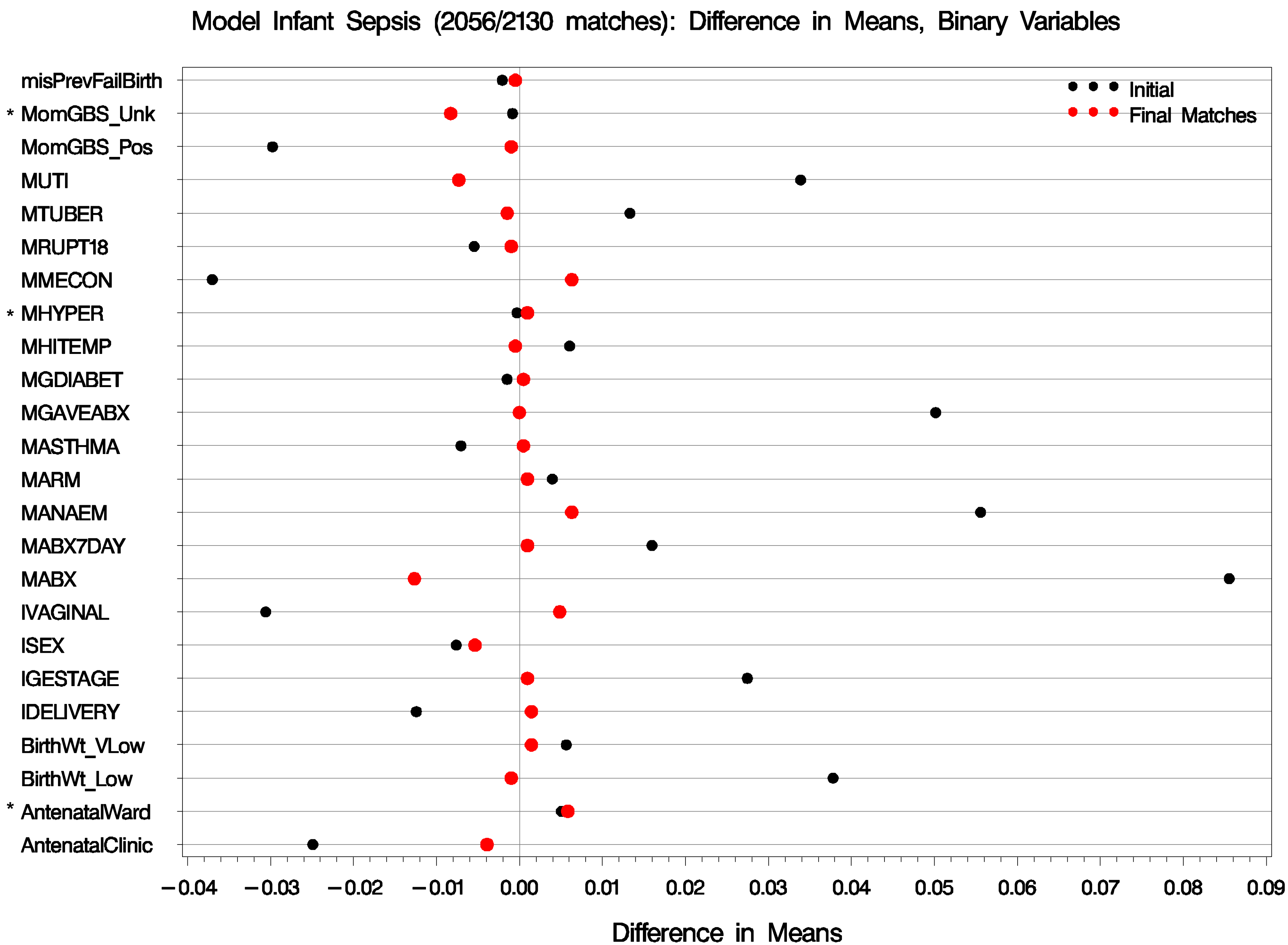
Marketing Example: Achieved Balance

- 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

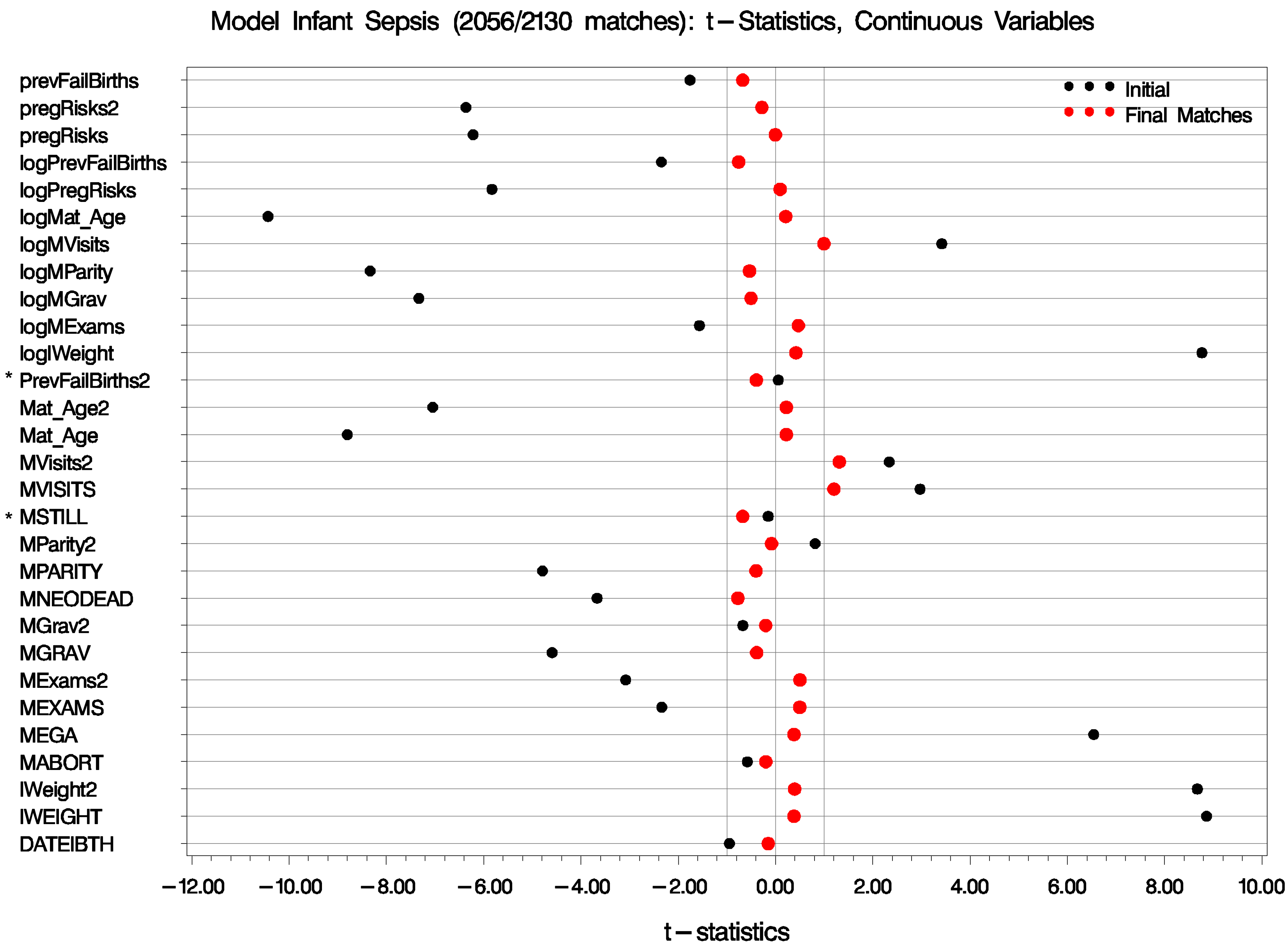

Matched Model—Duration 4+ hrs ExactMatch State/Term of Birth: Difference in Means, Binary Variables
Initial
Matched
uti
schl_csec
rom18_best
racebw
prevgbs
pngbstest
mr_gbsres
iptemp
chorio
Year2003
TN
OR
NY
NM
MN
MD
GA
CT
CO
CA
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Difference in Means

Matched Model—Duration 4+ hrs ExactMatch State/Term of Birth: t—Statistics, Continuous Variables





Model Infant Sepsis (2056/2130 matches): Difference in Means, Binary Variables
misPrevFailBirth
* MomGBS_Unk
MomGBS_Pos
MUTI
MTUBER
MRUPT18
MMECON
* MHYPER
MHITEMP
MGDIABET
MGAVEABX
MASTHMA
MARM
MANAEM
MABX7DAY
MABX
IVAGINAL
ISEX
IGESTAGE
IDELIVERY
BirthWt_VLow
BirthWt_Low
* AntenatalWard
AntenatalClinic
Initial
Final Matches
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
Difference in Means



Model Infant Sepsis (2056/2130 matches): t−Statistics, Continuous Variables
Initial
Final Matches
prevFailBirths
pregRisks2
pregRisks
logPrevFailBirths
logPregRisks
logMat_Age
logMVisits
logMParity
logMGrav
logMExams
logIWeight
* PrevFailBirths2
Mat_Age2
Mat_Age
MVisits2
MVISITS
* MSTILL
MParity2
MPARITY
MNEODEAD
MGrav2
MGRAV
MExams2
MEXAMS
MEGA
MABORT
IWeight2
IWEIGHT
DATEIBTH
−12.00 −10.00 −8.00 −6.00 −4.00 −2.00 0.00 2.00 4.00 6.00 8.00 10.00
t−statistics