APBG 2020 Webinar Series:
Comparative Effectiveness methods for Real World Data

July 2020

Alan J. M. Brnabic
Principal Research Scientist
Real World Analytics GSS
Chair, Australian Pharmaceutical Biostatistics Group

Acknowledgements:
Doug Faries, Tony Zagar, Zbig Kadziola, Mark Belger, Ilya Lipkovich

Myth Buster: Observational Studies (RWD)

- Simple studies to design
- Simple studies to implement
- Data quality is not important
- Analysis is relatively simple

Outline

- The Question
  - Real World = Comparing outcomes for 2 or more treatments
- The Problem
    - No randomisation-> Selection Bias
    - Confounding
    - Switching
    - Unmeasured confounding
- Approaches to adjust for confounding
  - Regression
  - Propensity score
  - Matching
  - Model Averaging

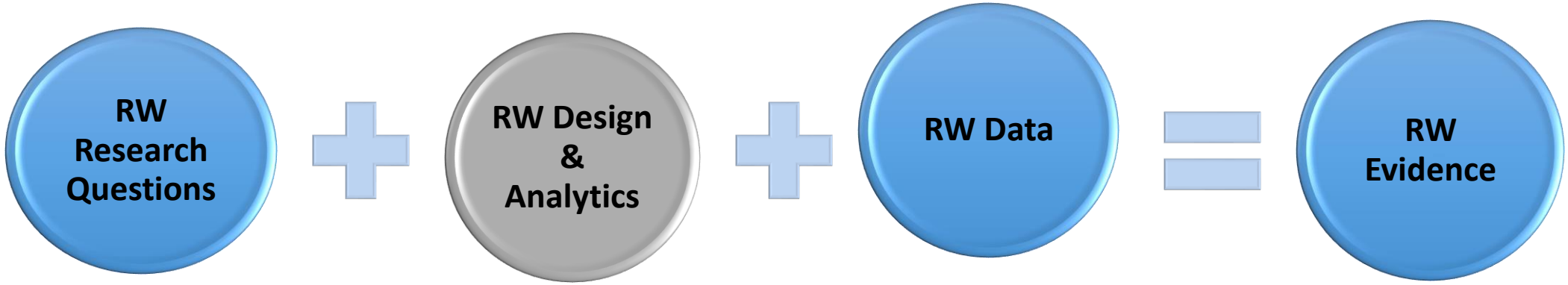
The Question: RWE Equation

Compare treatments

Selection Bias control methods

Any: Prospective/Retrospective

Comparative effectiveness

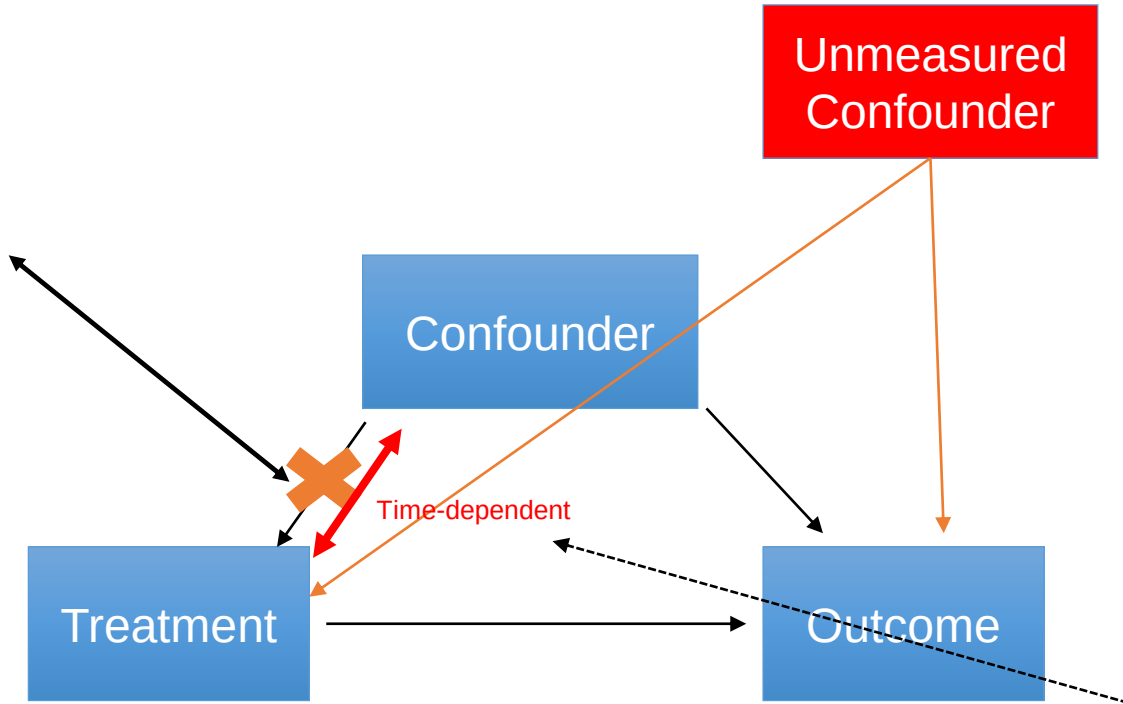


RWD is necessary but not sufficient for generating RWE

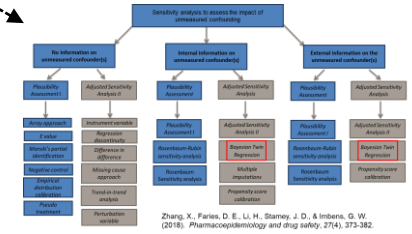
Duke-Margolis Whitepaper: A Framework for Regulatory Use of Real World Evidence; Sept 13, 2017

The Problem: Summary

Inverse Propensity Weighting	Regression
Propensity stratification	Propensity matching



UC Sensitivity



Zhang, X., Faries, D. E., Li, H., Stamey, J. D., & Imbens, G. W. (2018). *Pharmacoepidemiology and drug safety*, *27*(4), 373-382.

Unlike RCT, treatment assignments in real world observational studies are not randomized but usually influenced by confounders prior to treatment initiation, therefore estimated causal treatment effect could be biased without proper confounder adjustment.

G-methods

Education Corner

An introduction to g methods

Ashley I Naimi[1]*, Stephen R Cole[2] and Edward H Kennedy[3]

[1]Department of Epidemiology, University of Pittsburgh, [2]Department of Epidemiology, University of North Carolina at Chapel Hill and [3]Department of Statistics, Carnegie Mellon University

The Problem: Sources of Bias

- Information/Measurement Bias
- Information for patients in Groups A and B are gathered or measured differently
- Unblinded (perhaps blind raters), case control
- Sponsor Bias
- Conduct differs due to knowledge of sponsor
- Consider PRO, retrospective data

- Selection Bias
- Groups A and B differ in some important aspect other than treatment
- Imbalance in patient characteristics between groups is typical
  - medication choices in the real world are not made at random, but on a set of perhaps complex factors
- Confounding variables
  - associated with both treatment selection and outcome

The Problem: Selection Bias

With randomization – standard methods produce estimates of causal treatment effects

Selection

Without randomization – standard methods produce only ‘associations’ …. Treatment groups are NOT comparable at baseline thus comparisons are BIASED

Issue:

Confounding

Patients can switch between treatments and take multiple combinations of treatments throughout the trial period

Bias

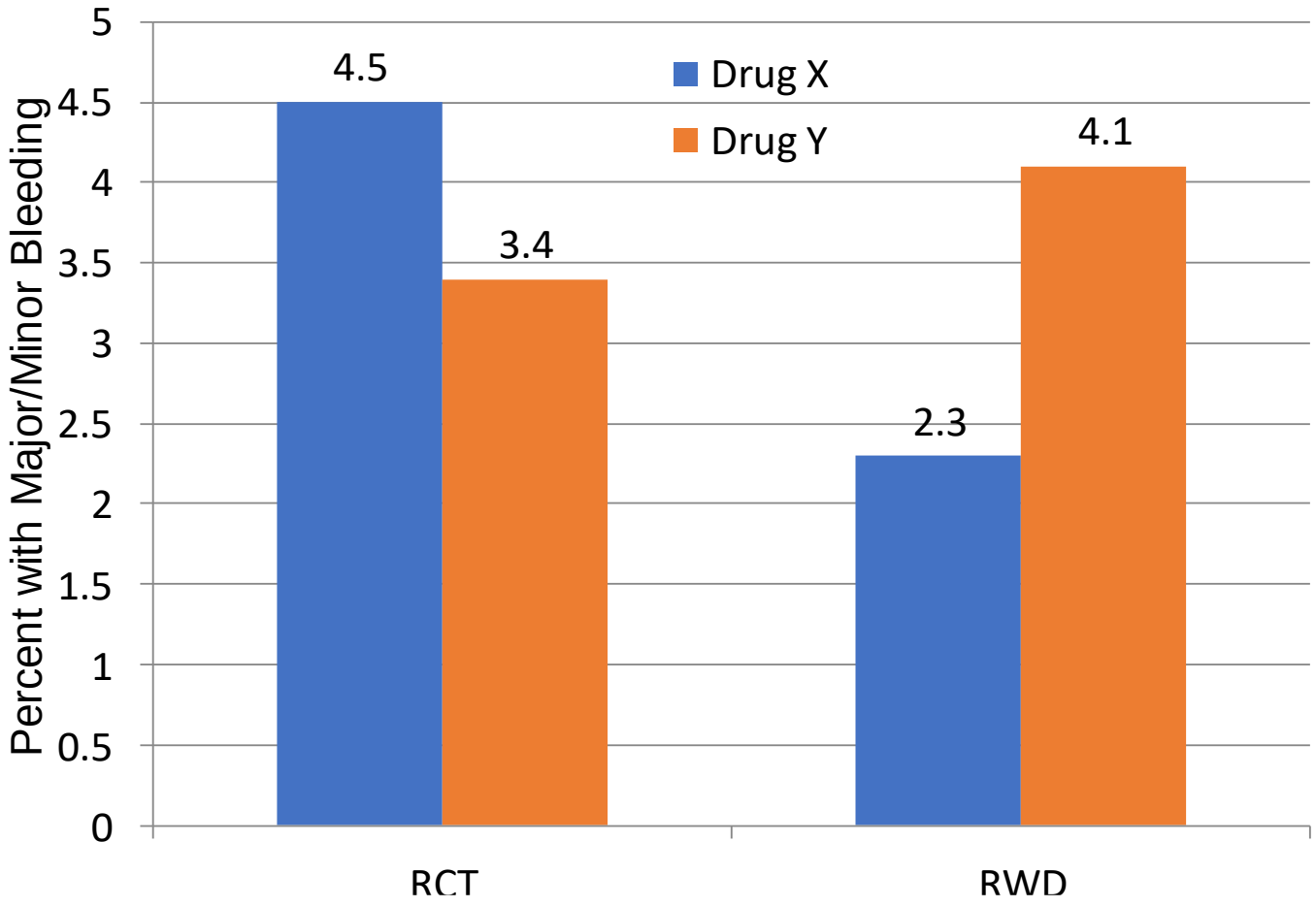
Complex data analysis ->quality of the results relies on quality of the data

The Problem: Confounding (measured)

- In Comparative effectiveness research patient characteristics may differ between treatment groups.
- Unadjusted results tell you the differences between treatment management decisions
- They do not tell you about relative effectiveness of two treatments

The Problem: Confounding (measured)

Bleeding Rates in ACS-PCI* patients RCT vs RWD



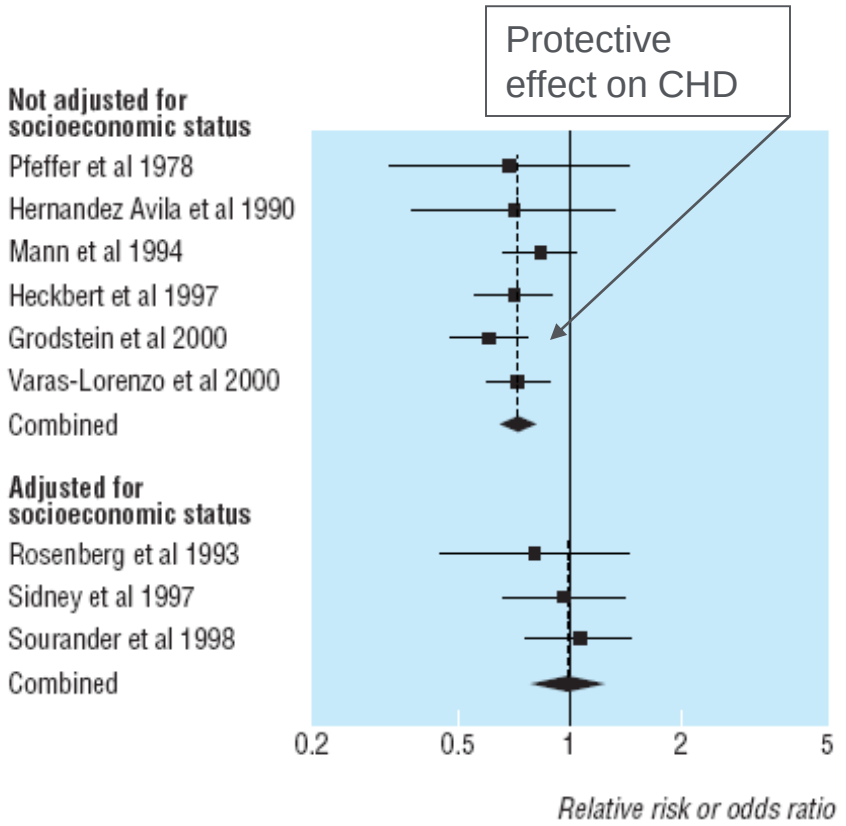
- Due to Design: Bleeding rates will be lower in patients receiving Drug X (as given to patients with a low risk of bleeding)
- This study demonstrates 2 Distinct populations = nonoverlapping distributions
- Message: Management of patients is effective ➲right treatment to right patients
- Question asked may mean you do not want to adjust

Premier Database	Drug X (N=9,404)	Drug Y (N=74,163)
Mean Age (years)	56.4	60.4
Age > 75 yrs (%)	**1.1%**	**9.1%**
Male (%)	76.9%	69.5%
Renal Insufficiency	7.6%	13.5%

*Percutaneous coronary intervention (PCI) Acute coronary syndrome (ACS)

The Problem: Confounding (measured)

- Higher socioeconomic position is strongly associated with:
  - more frequent use of HRT
  - lower risk of CHD
- No effect if adjusted for socioeconomic status
- Note: In the large Women's Health Initiative RCT:
  - HRT had **no** beneficial effect on CHD



Meta-analysis of cohort studies and case-control studies of hormone replacement therapy and coronary heart disease. There is little evidence for a protective effect when analyses are adjusted for, in contrast to studies not adjusted for, socioeconomic status. Adapted from Humphrey et al, reference 7

Erik von Elm, Matthias Egger The scandal of poor epidemiological research Reporting guidelines are needed for observational epidemiology BMJ VOLUME 329 16 OCTOBER 2004

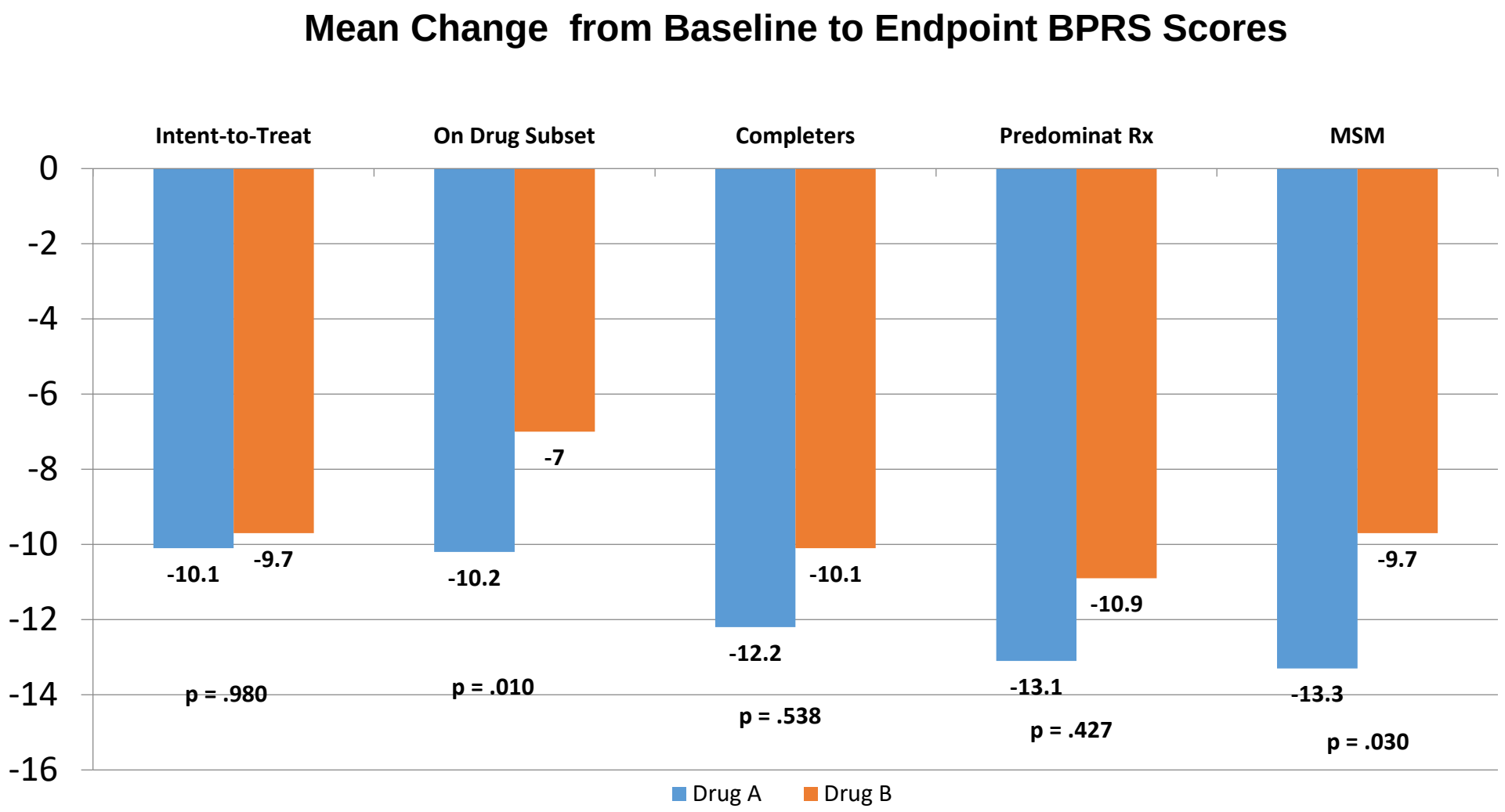
The Problem: Analysis populations

- **Different populations** used in observational studies for same analysis answer **different questions**:
  - Completers
  - Monotherapy
  - Remain on monotherapy
  - Intention to treat
  - Observed treatment
  - Predominant treatment e.g Defined Daily Dose

Think carefully about which population best answers the research question.

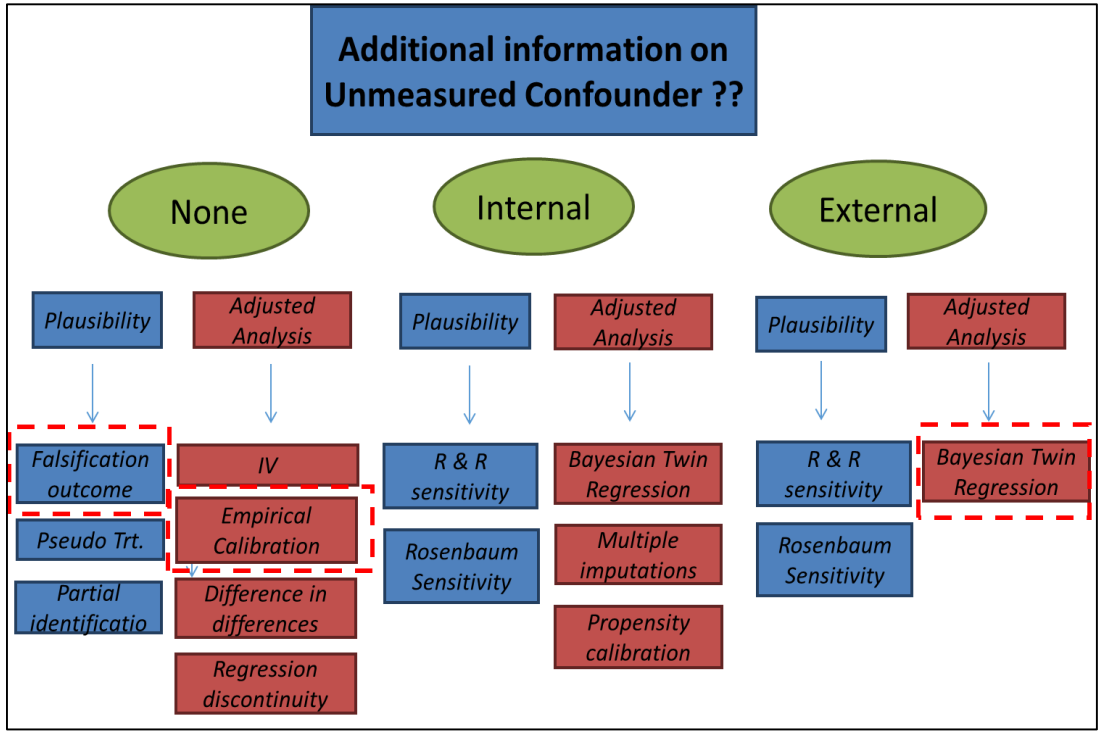
For example what happens if there is switching?

The Problem: Analysis populations (switching)



The Problem: Unmeasured confounding

- Collect as much data on potential confounders as possible
  - Do your research up front
- Assess sensitivity
  - Instrumental variables
  - Sensitivity – level of unmeasured confounding necessary to change inferences
  - Internal/external sensitivity (using additional data sources)



VALUE IN HEALTH 16 (2013) 259–266

Available online at www.sciencedirect.com
SciVerse ScienceDirect
journal homepage: www.elsevier.com/locate/jval

Evaluating the Impact of Unmeasured Confounding with Internal Validation Data: An Example Cost Evaluation in Type 2 Diabetes

Douglas Faries, PhD[1,*], Xiaomei Peng, MS[1], Manjiri Pawaskar, PhD[1], Karen Price, PhD[1], James D. Stamey, PhD[2], John W. Seaman Jr., PhD[2]

[1]Eli Lilly & Company, Indianapolis, IN, USA; [2]Department of Statistics, Baylor University, Waco, TX, USA

PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2016
Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/pds.4053

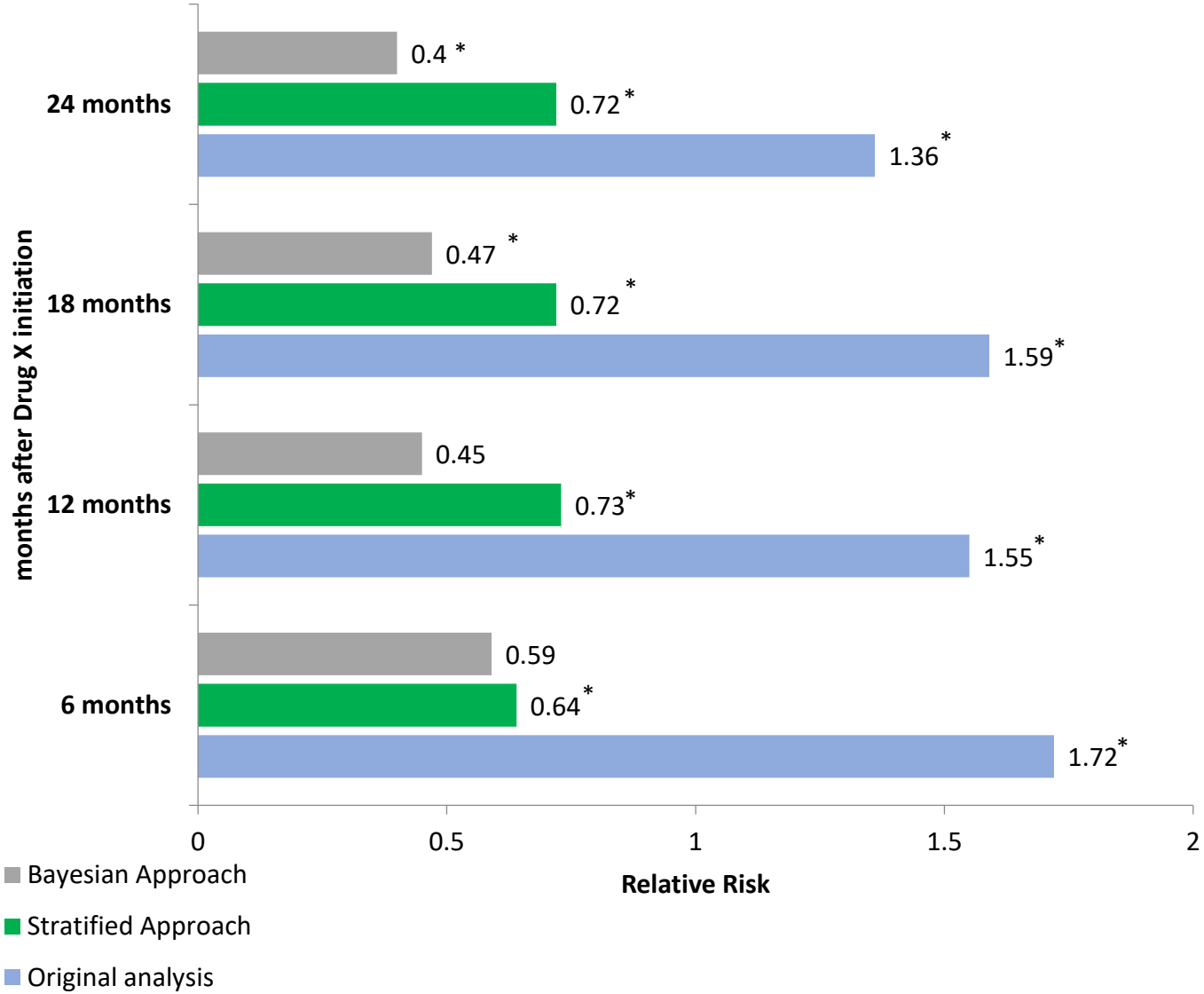
ORIGINAL REPORT

"A Bayesian sensitivity analysis to evaluate the impact of unmeasured confounding with external data: a real world comparative effectiveness study in osteoporosis"

Xiang Zhang[1]*, Douglas E. Faries[1], Natalie Boytsov[1], James D. Stamey[2] and John W. Seaman, Jr.[2]

[1]*Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN, United States*
[2]*Department of Statistical Science, Baylor University, Waco, TX, United States*

The Problem: Unmeasured confounding



* Statistically significant at p<0.05 level.

- In RCTs, *DRUG X* was shown to be more effective in reducing the risk of fractures compared to non-active patients
- In real-world setting, the analysis without BMD failed to show any effect of *DRUG X*
- Adjusting for BMD (Controlling for unmeasured confounding) reversed the negative findings and changed them to showing a significant positive effect for *DRUG X*

The Problem: Bias Adjustment Quote

D'Agostino (JAMA 2007)

"… even with the best of designs, observational studies, unlike the RCTs, do not automatically control for selection biases. Therefore, statistical methods involving matching, stratification, and/or covariance adjustment are needed"

Bias Adjustment Tools

REGRESSION METHODS

Generalized Linear models (e.g. Logistic regression)

Penalised regression models

Longitudinal models

PROPENSITY SCORE METHODS

Regression with weights/strata*

Matching (with/out PS)

Stratification and Clustering

Inverse weighting

NEW: Tree based methods

LONGITUDINAL METHODS

Epoch analysis

Marginal structural modelling

Structural nested modelling

Propensity score regression

~G-methods

NEW: Prognostic Score

INSTRUMENTAL VARIABLES

NEW: MODEL AVERAGING**

(applying machine learning approaches)

*Doubly robust methods

**Zagar, Kadziola, Lipkovich, Madigan, Faries: submitted 2020

Methods: Regression

- Historically the most common approach to adjust for confounders
- **Traditional Examples:**
  - Multivariable Regression- continuous outcomes
  - Logistic regression models – binary outcomes
  - Poisson Regression – Event rates outcome
  - Cox Proportional Hazard's models – Time to event outcome
- The models fit treatment alongside important patient baseline characteristics.
- Results give treatment effect after adjusting for all other factors in the model.
- Newer regression methods include penalized regression (PR):
  - Traditional GLM cannot manage large number of main effects (interactions) – especially when this > sample size
  - Methods: Ridge, Lasso, Elastic Net
  - 2 step process (replaces backward selection when you have many terms -> overfit)
    - Use PR to select covariates
    - Then fit model in via GLM (to get statistics) based on PR results

Methods: Traditional Regression

- Regression is not one of the recommended approaches
- There is good literature that shows most of the time the results of regression do not differ from propensity scoring
- However, there is good literature showing it can be biased when there are larger differences between groups
- Regression fails when there is little overlap between the cohorts – and is difficult to assess

Methods: Matching

- Number of different approaches
- No Gold Standard Recommended
- Simple: case-control matching on one factor (EXACT)
- Complex: Entropy Balancing method
- Matching algorithms include
  - Greedy nearest neighbour (w/o replacement only)
  - Optimal matching (w/o replacement only)
    - Fixed ratio matching,
    - variable ratio matching
    - full matching
  - Matching with replacement
  - Genetic matching
- Can be problem if you have low sample size

Methods: Propensity Score (PS)

- For 2 treatments: Usually generated with a logistic regression model to identify factors associated with choice of treatment.
- Model gives a probability (PS) on the likelihood of an individual patient receiving treatment A vs treatment B.
- Propensity scores are then used to balance treatment effect

- For >2 treatments Generalised PS
  - GPS matching
  - Inverse Probability Weighting
  - Vector Matching
- Other approaches to creating PS:
  - Tree methods: Gradient boosting
  - 2 step: Penalised regression then generalised linear models
  - More later

Methods: Propensity Score (PS)

Matching

Combine matching algorithms. Match patients exactly (age gender) then with similar PS, then compare Cohorts of matched pairs

Stratification

Group patients with similar PS; Compare cohorts within each PS strata; then average across the strata

Regression

Simple regression model

Y = Treatment + PS

Inverse Weighting

Run weighted analysis, weighting each patient by the inverse of P(Being on actual Treatment)

Methods: Which to choose?

- Comparative Effectiveness estimates are sensitive to the Analytic model

Volume 12 • Number 8 • 2009
VALUE IN HEALTH

Good Research Practices for Comparative Effectiveness Research: Analytic Methods to Improve Causal Inference from Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III

Michael L. Johnson, PhD,[1] William Crown, PhD,[2] Bradley C. Martin, PharmD, PhD,[3] Colin R. Dormuth, MA, ScD,[4] Uwe Siebert, MD, MPH, MSc, ScD[5]

ISPOR highlighted the issue (2009,2012) but gave no guidance of what to do – except be transparent about what you do ☺

Value Health. 2012 Mar-Apr;15(2):217-30. doi: 10.1016/j.jval.2011.12.010.

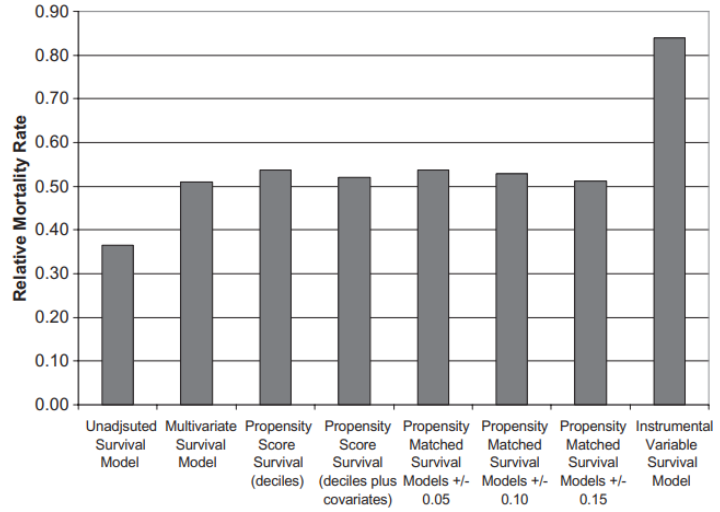
Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report.

Berger ML[1], Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL.

JAMA. 2007 Jan 17;297(3):278-85.

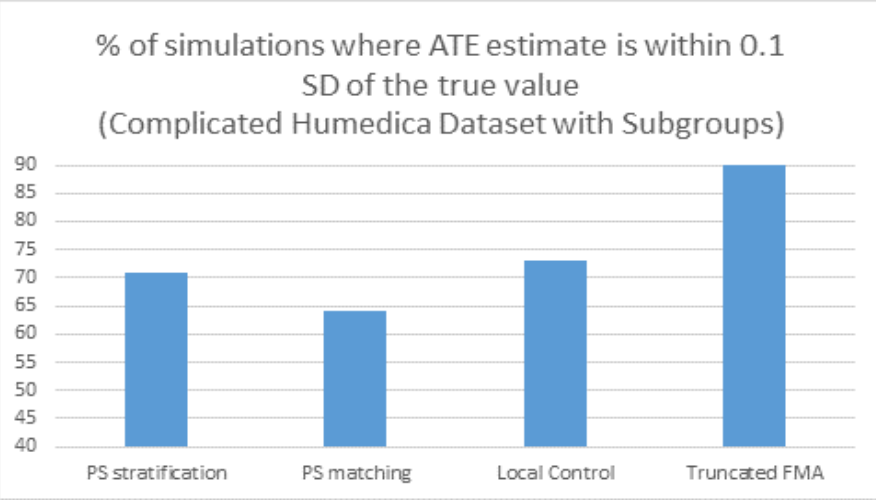
Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods.

Stukel TA[1], Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ.

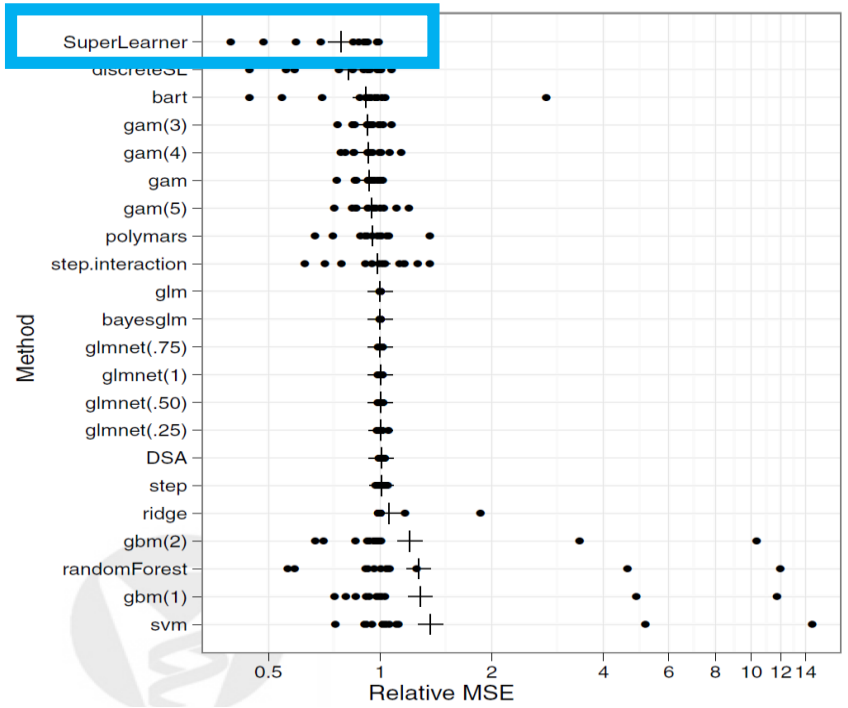


Methods: Which to choose?

- In the causal inference space
  - Simulations show we are 10% right more often when use machine learning/data driven ensemble methods (Zagar et al: submitted 2020)



- In the Predictive analytics space
  - Applying multiple models is becoming popular (Hess, Brnabic – work in progress)
  - Super Learner research shows better operating characteristics (Polley and Van der Laan 2011



Methods: Which to choose?

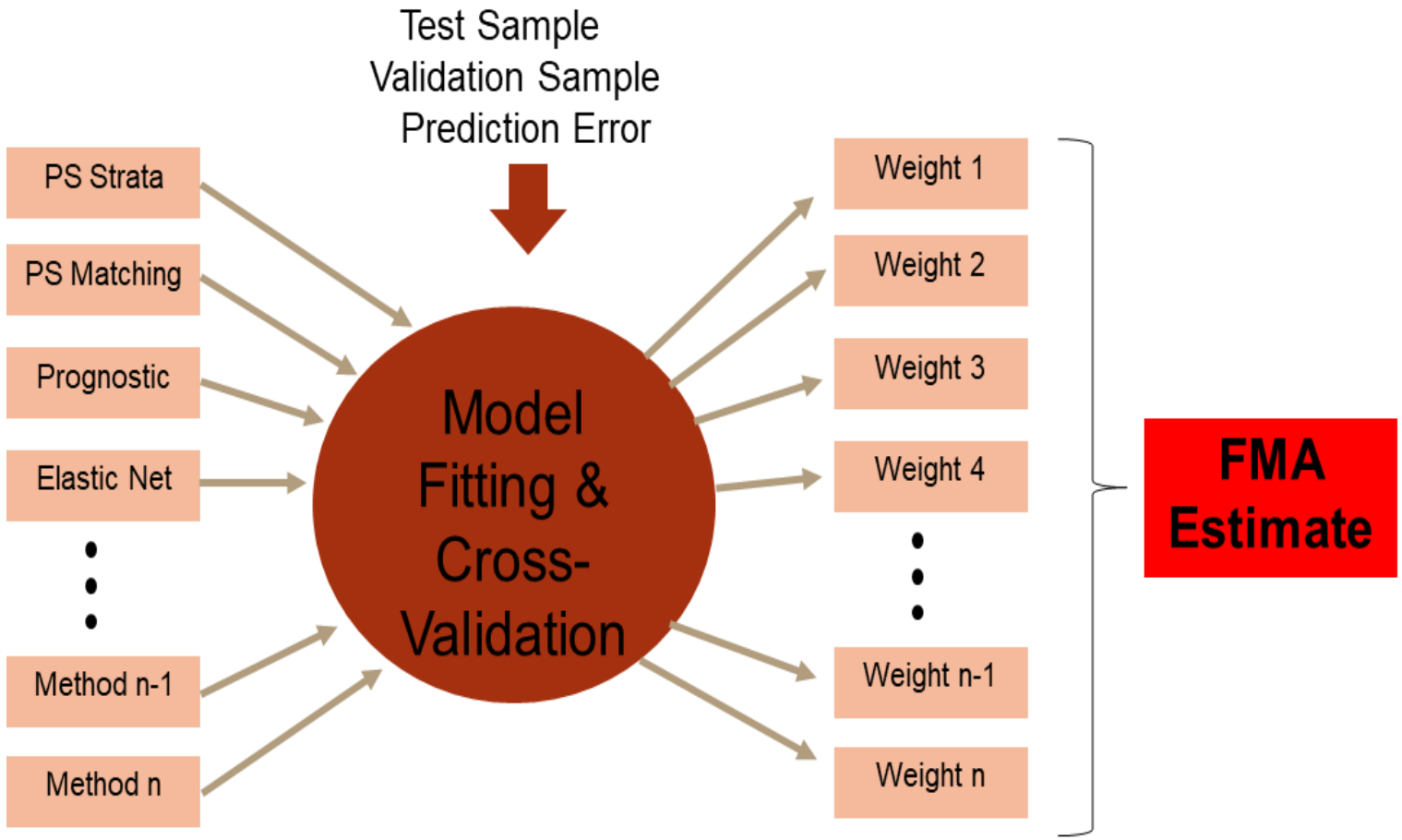
- Older approaches – Regression methods (linear, logistic, Cox models)
- More recently, Propensity score methods have become the ***gold standard***
- Currently, at least another analysis different to the primary analysis is recommended (sensitivity analysis)
- Future state: Model averaging Approaches
  - looks at many methods – also known as ensemble methods ,super learner,
  - uses machine learning principles

Methods: The future->Model Averaging

- Has been around for several decades just recently applied to causal inference
- Estimates weights for each potential analytical approach
- Final estimate is a weighted average of estimates across all the methods – where the weight applied to each method is based on the amount of evidence supporting the method
- Used directly to estimate the treatment effect rather than the propensity model (Zagar et al: submitted 2020)

Methods: The future->Model Averaging

- We believe the field is moving from 1 favourite method to using Machine learning/Data driven algorithms to choose the best
- Could be Bayesian or Frequentist
- Weighted average of all models tested
- Treatment effect estimates weighted based on k-fold cross validation
- Currently time consuming & complicated



Zagar, Kadziola, Lipkovich, Madigan, Faries: submitted 2020

Steps: Model Averaging

1. Selection of Individual Methods
2. Computing MA Weights based on mean square predicted error (MSPE) from cross validation. Added complexities inherent due to the fact that not all causal methods produce a predicted value for each individual. Modified formula:

$$W = \exp\left(-\frac{\widehat{MSE}_{CV}}{\hat{\sigma}^2}\right)$$

where the variance is the estimated pooled within treatment variance using the full (both training and hold out) data. The mean squared prediction error for an individual method is then computed as follows across all N patients.

$$\widehat{MSE}_{CV} = \mathrm{N}^{-1}\left[\sum_{i=1}^{N_0}(Y_i - \tilde{Y}^{-D(i)}(X = x_i, T = 0))^2 + \sum_{i=1+N_0}^{N}(Y_i - \tilde{Y}^{-D(i)}(X = x_i, T = 1))^2\right]$$

3. Calculate MA Estimator:

Based on M analytical methods

$$\hat{\delta}_{FMA} = \frac{\sum_{m=1}^{M}\hat{\delta}_m W_m}{\sum_{m=1}^{M} W_m}$$

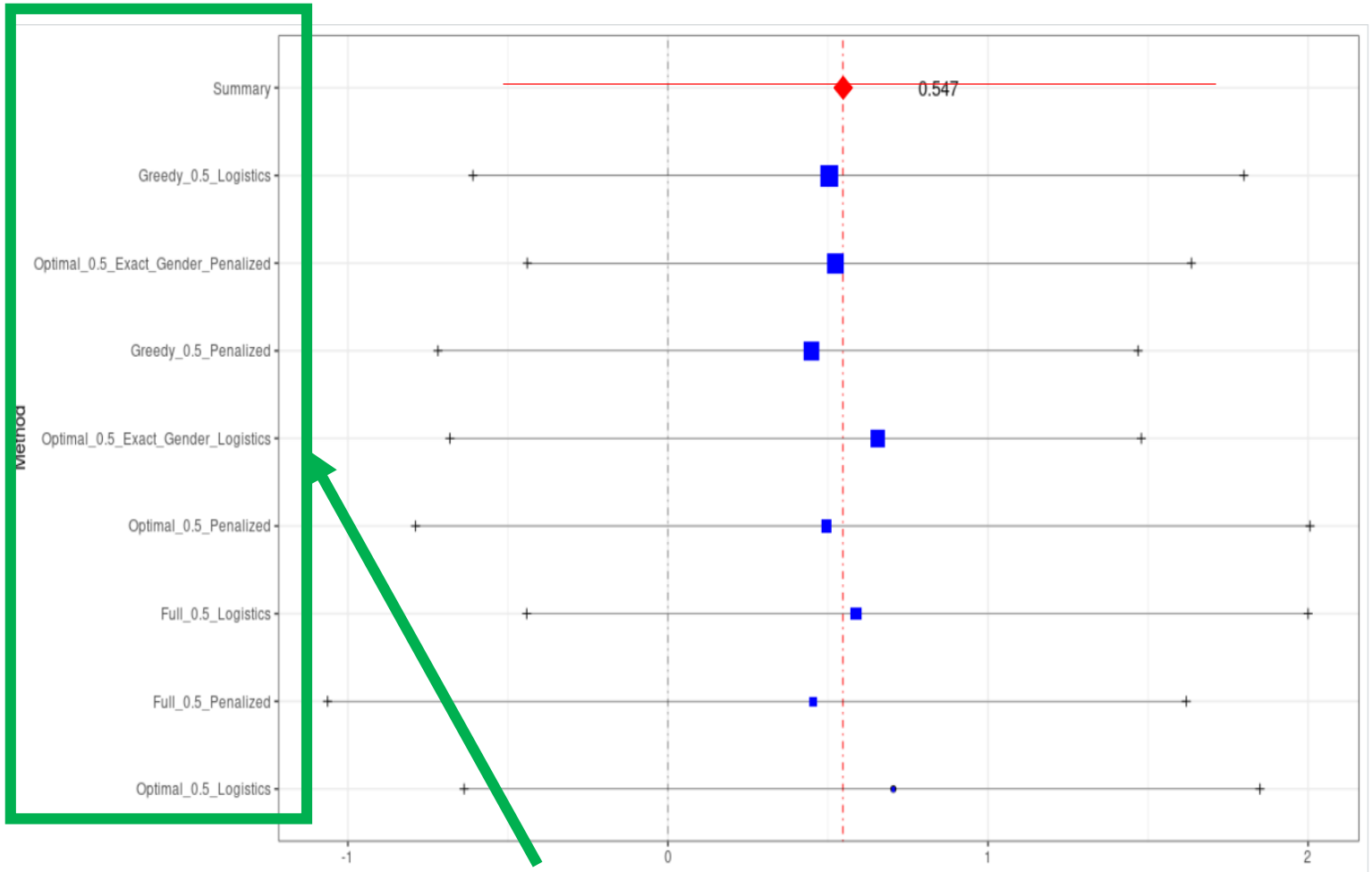
Bootstrapping using the percentile method is used to estimate the variance of the model averaging estimator and allow inferences of the results

Methods: Frequentist Model Averaging

- Weighted average of all models fitted displayed at the top
- CIs are bootstrapped
- Treatment effect estimates weighted based on influence
- Methods are ordered from:
  - highest weight (most influential) to
  - lowest weight (least influential)

FMA Analysis (Matching methods)

Forest Plot



Different combinations of matching methods

The Problem: Assumptions for causal inference

Propensity Score adjustments can provide for estimates of the causal treatment differences under the following assumptions:

#1 **No Unmeasured Confounders**

All confounders are in the dataset and analysis

#2 **Sufficient Overlap in Populations**

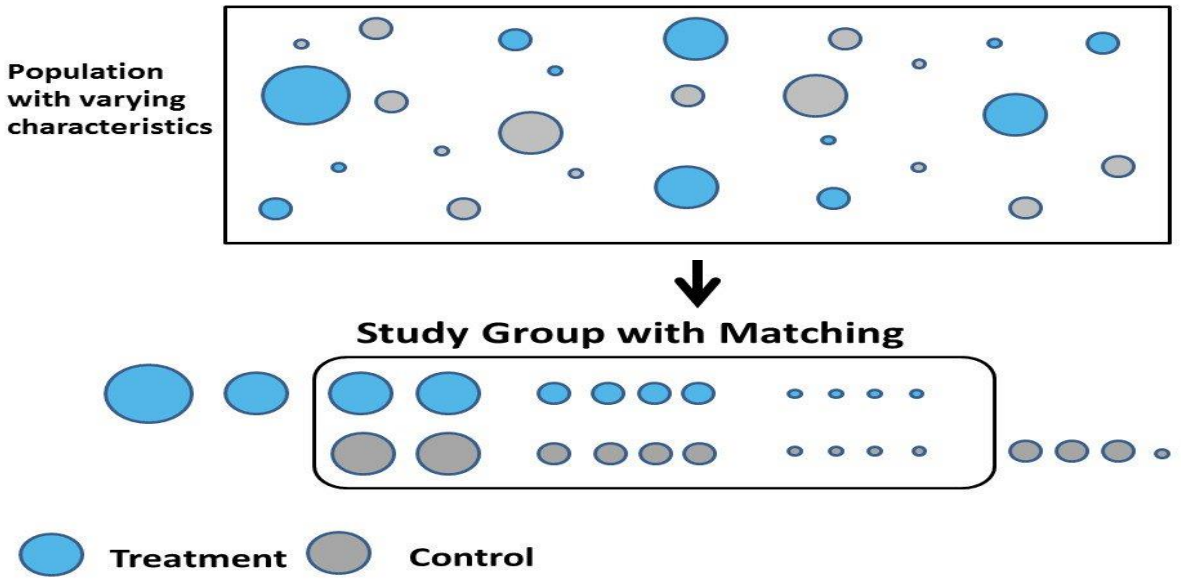
Positivity, no perfect confounding e.g. cannot have all females in one group

#3 **Correct Statistical Models**

Propensity Scores

- Want to know the effect of something comparing two or more treatments
- You **don't** have random assignment
- You **do** have a list of variables that determine whether or not an individual received treatment
- Used to minimise selection bias by balancing covariates (the characteristics of participants) between treated and control groups
- When the covariates are balanced, it becomes much easier to match participants and assess causality

- The probability that a unit with certain characteristics will be assigned to the treatment group - as opposed to the control group



Patients are similar in some way =Probability (assigned to a treatment)

Example: Current recommended approach to PS

1. Select Covariates for Adjustment
2. Compute Propensity Score
3. Assess Feasibility & Confirm Balance – " Outcome free"
4. Match Treated and Controls based on Algorithm (e.g. 1:1 Greedy) or PS regression or PS stratification
5. Analysis: TTest/Regression on Matched Sample

*Ho et. al, 2007

Rubin – Done without looking at the outcome data-> not biased with influence of the outcome

Example: Current recommended approach to PS

1. Select Covariates for Adjustment

- In general, there are 3 sets of covariates we may consider for inclusion in the estimation model:

a) Covariates that are predictive of treatment assignment

b) Covariates that are associated with the outcome variable

c) Covariates that are predictive of both treatment assignment and the outcome

Example: Current recommended approach to PS

2. Compute Propensity Score

1. A priori logistic regression model – prespecified list of covariates
   - In SAS: PROC PSMATCH or PROC LOGISTIC

2. Automatic parametric model selection – Iterative process - examines all ME & interactions and assesses balance across strata then select model based on terms that improve the balance the most

3. Two stage approach: use penalized regression then GLM

4. Nonparametric: can handle missing without imputation & sensitive to outliers e.g CART, bagged CART, Random forests, boosted CART (gradient boosting)

- In SAS: PROC GRADBOOST (xgboost in R)
- *Has nice autotune function for tuning hyperparameters (SAS VIYA)*

☆Select based on the Quality of the PS estimates = good balance between comparison groups

Example: Current recommended approach to PS

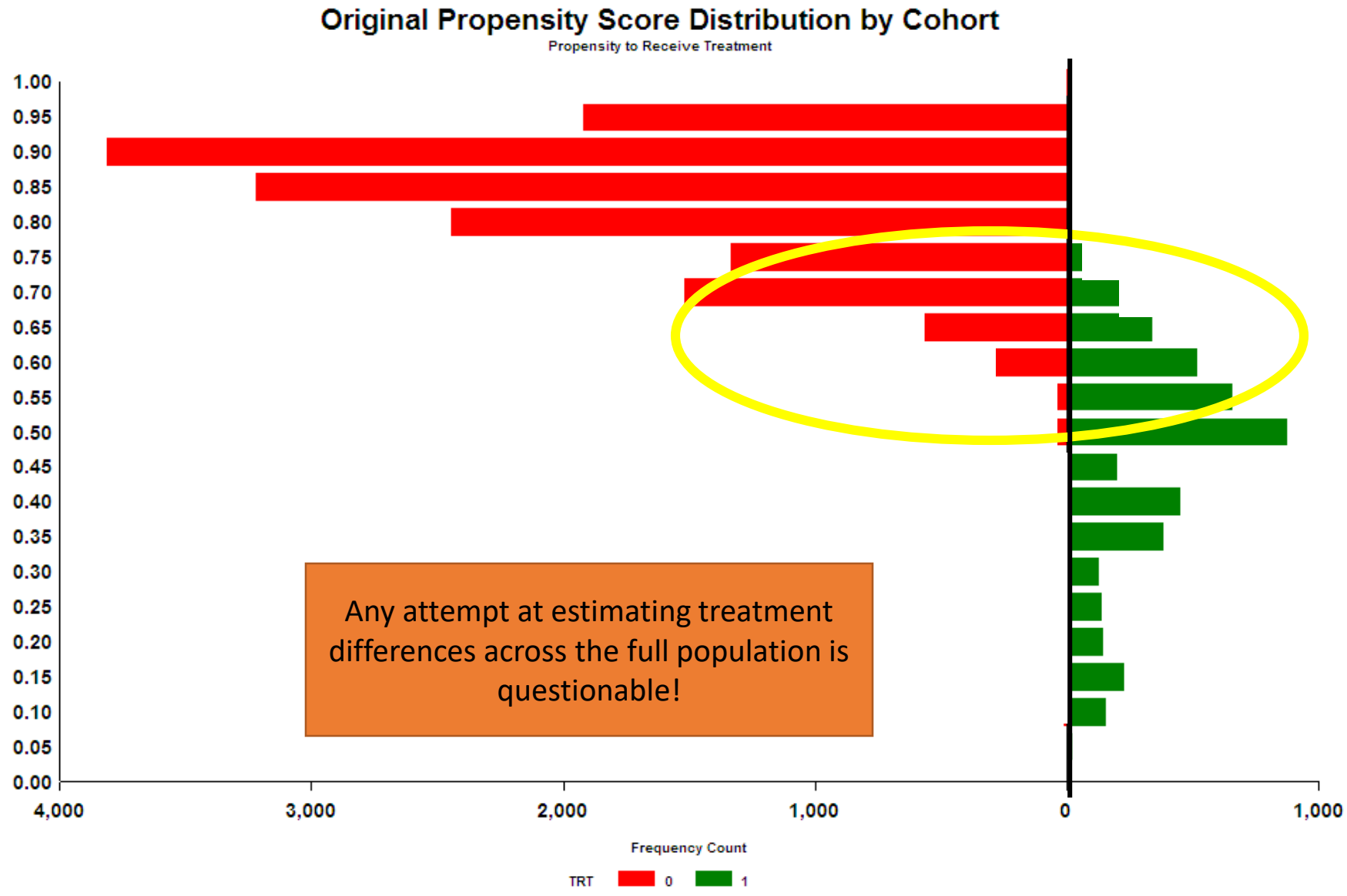
3. Assess Feasibility & Confirm Balance – "Outcome free"

- Confirming that the target population of inference is feasible with the current data ➲ Sufficient overlap to generalise
- Assessing the ability to address confounders (measured and unmeasured) ➲ must balance the two treatment groups in regards to all key covariates that may be related to both outcome and the treatment selection

Ensure the model adequately balances the covariates

"the success of the propensity score modeling is judged by whether balance on pretreatment characteristics is achieved between the treatment and control groups …" (D'Agostino 2007)

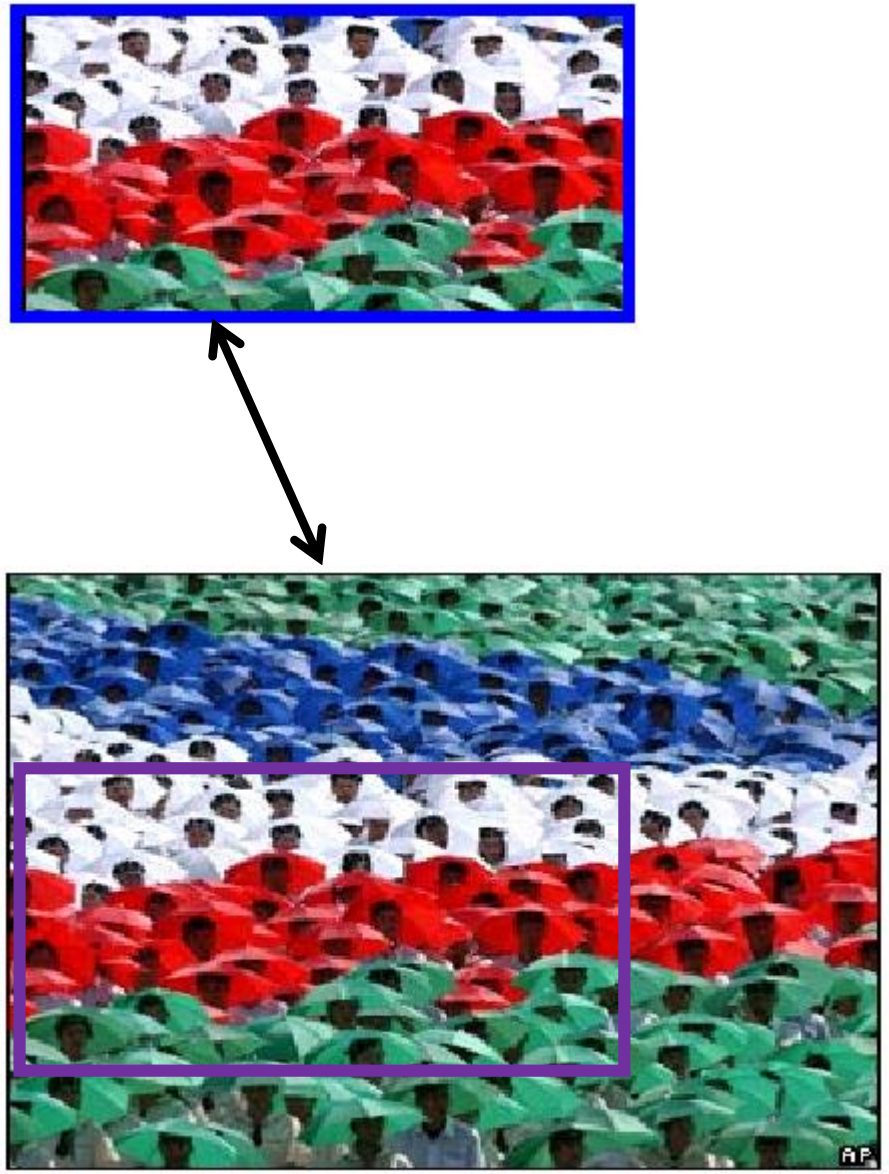
Assessing Propensity Distribution Overlap



Propensity Score: Interpretation

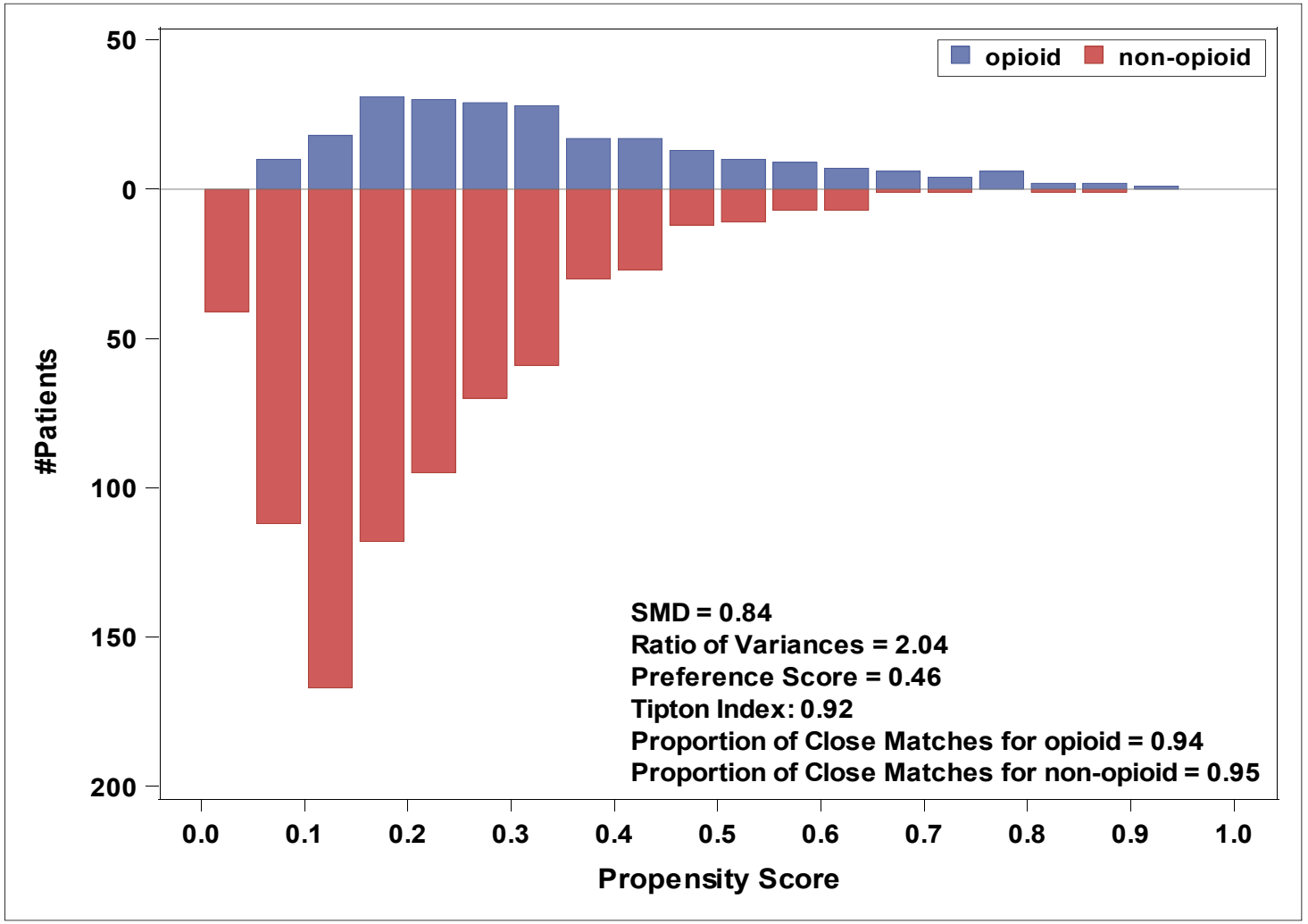
Matched patients:

- Now that we have a *different* group of patients with the 'new control group' we cannot generalise to the entire sample we had
- We are not calculating an average treatment effect but the treatment effect on the treated
- Results generalise to what *look like the treatment group* not the entire population



Assessing Propensity Distribution Overlap

Pre-matching/Pre-trimming



*SMD=Standardised mean difference

Feasibility Statistics for PS

SMD: <0.25 (Austin:<0.1) - OK

Ratio of variances: 0.5 to 2.0 - OK

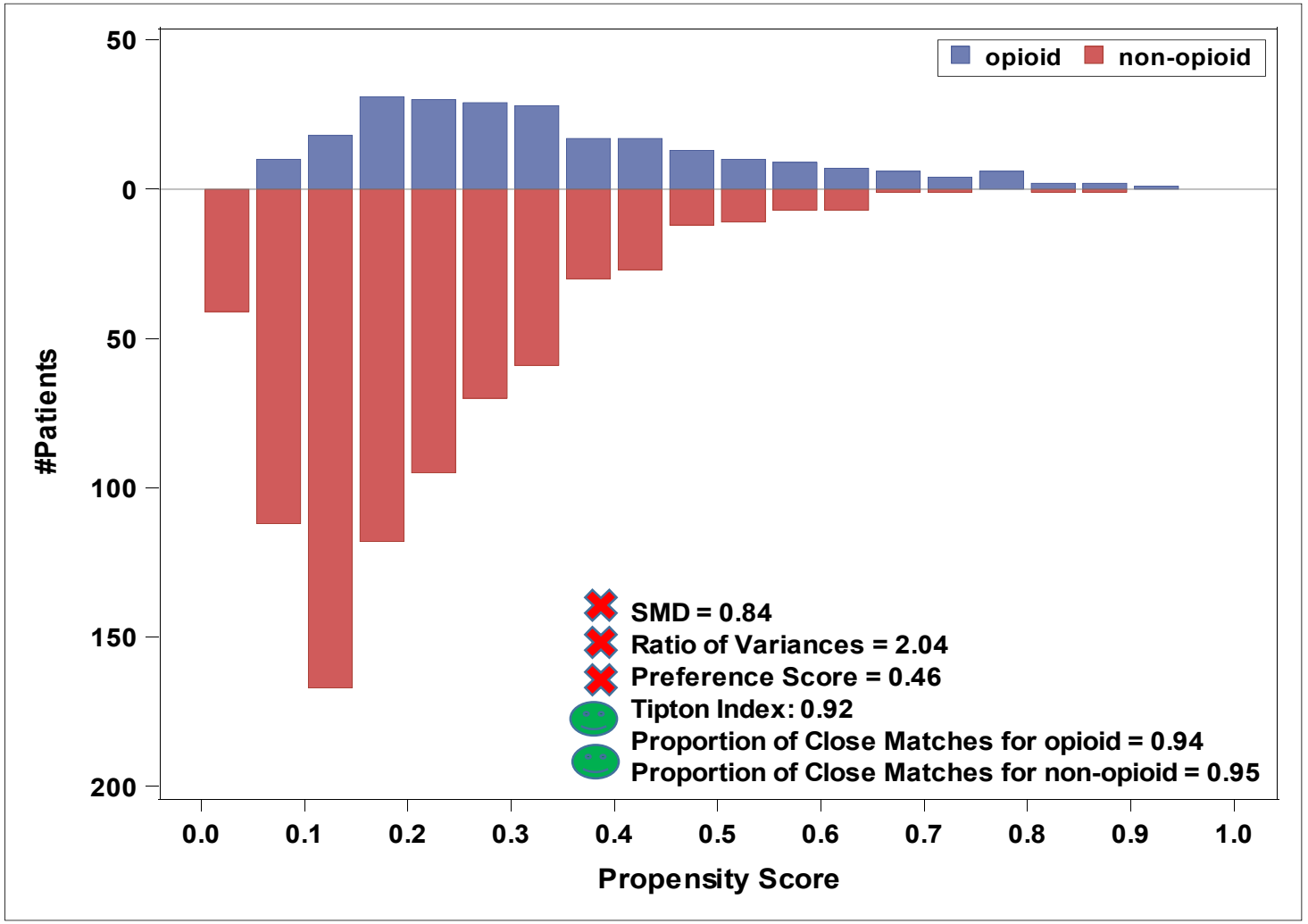
Preference score: 0.3 to 0.7 - OK

Tipton: >0.9 - OK

Proportion of close matches 1.0 = perfect (based on Greedy 1:1 with replacement)

Assessing Propensity Distribution Overlap

Pre-matching/Pre-trimming

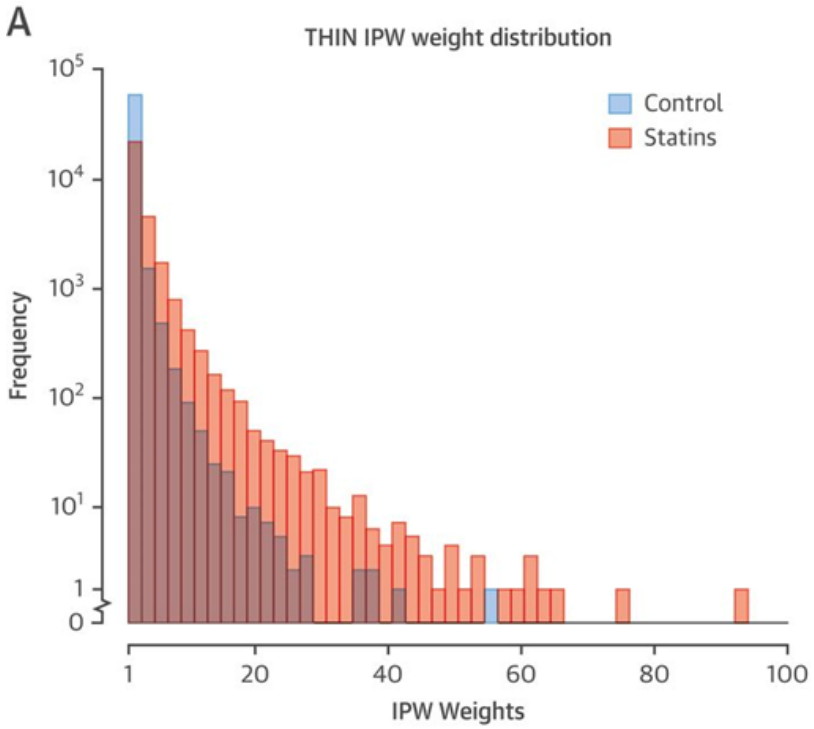


Feasibility Statistics for PS

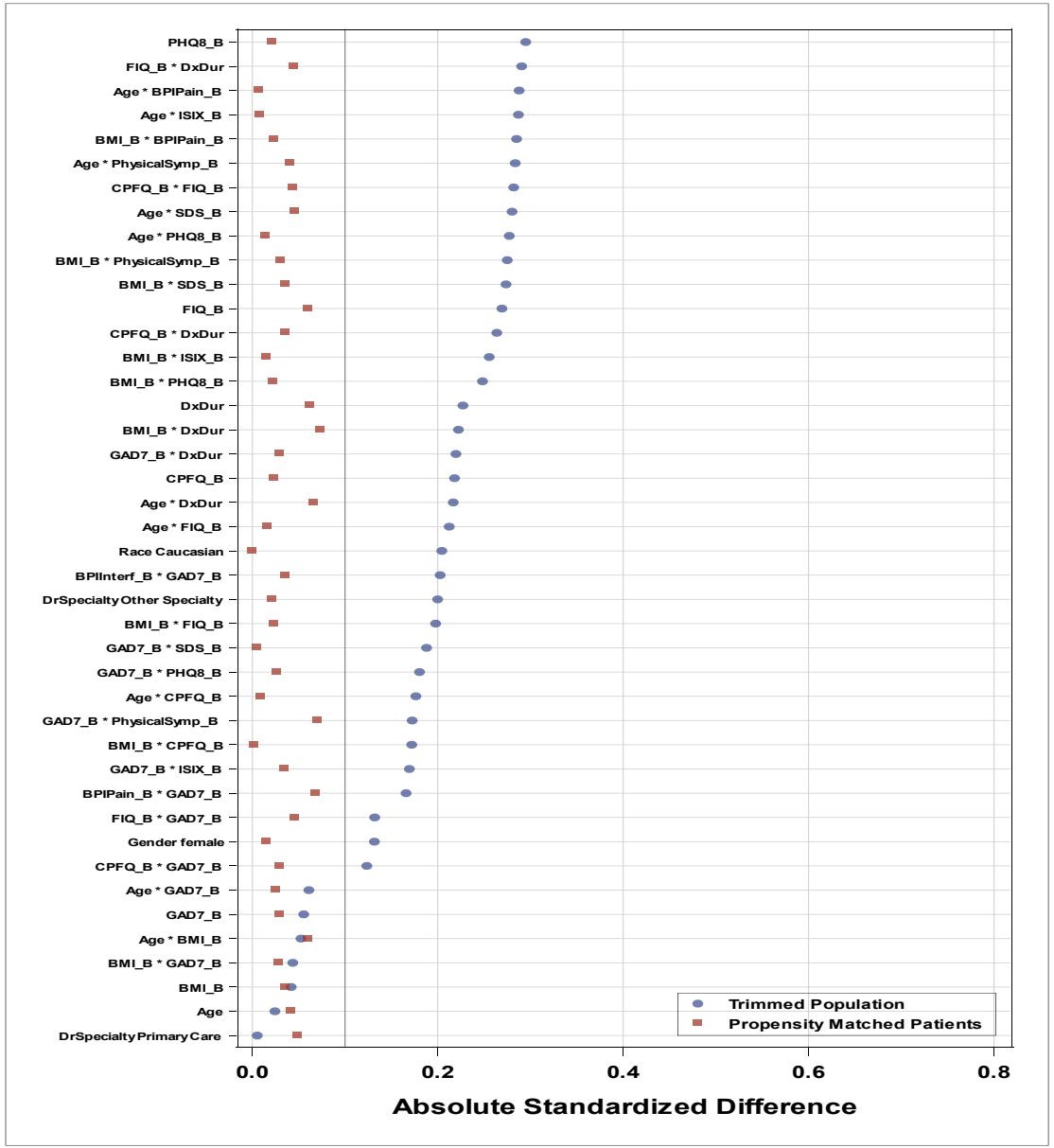
- Comparative analyses may be possible (similar ranges in the distributions)
- The success of the statistical adjustment, the assumption of unmeasured confounders, and some consideration of amending the target population is warranted.
- Standard matching will likely lead to a population of inference much closer to the Opioid population than the full population
- Methods such as matching with replacement or inverse weighting may rely on extreme weights to produce the balance necessary

Assessing Propensity Balance: Covariates

- How do we know how well matching worked- Assess covariate balance
- Use plots
  - Standardise mean difference (SMD) (<0.1)
  - Variance Ratios (VR) (has to be close to 1)
- Use permutation to get CI for SMD, VR
- Single Measures Comparing PS
  - Average Absolute SMD
  - Maximum Absolute SMD
  - % ASMD<0.1
- Trimming needed for PS weighting



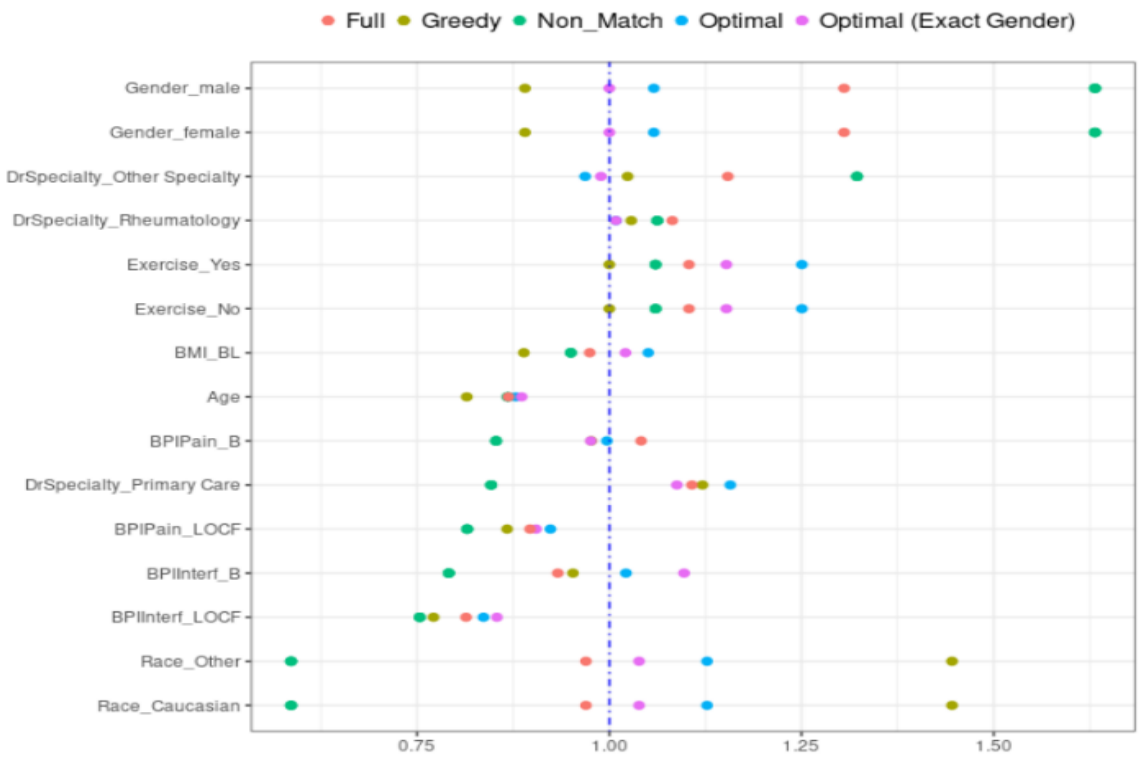
Assessing Propensity Balance: Covariates



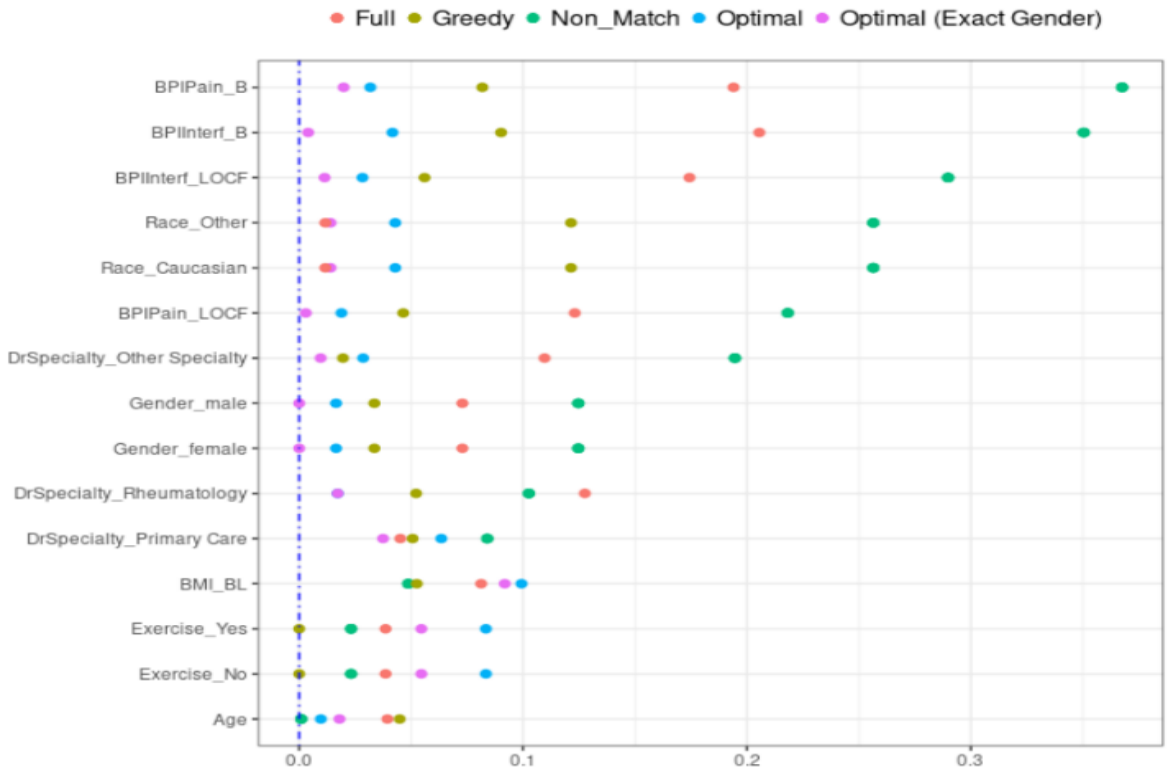
	CS Patients				Propensity Matched Patients					
Covariate	opioid (N=237)	non-opioid (N=715)	Std. Diff	Variance Ratio	opioid (N=224)	non-opioid (N=224)	Std. Diff	95% CI of Std Diff under H0: Std Diff=0	Variance Ratio	95% CI of Variance Ratio under H0: Variance Ratio=1
Age	50.3 (±11.28)	50.02 (±11.56)	0.02	0.95	50.17 (±11.22)	49.69 (±11.2)	0.04	(-0.18,0.18)	1.00	(0.71,1.29)
Gender female	214 (90.3%)	671 (93.8%)	-0.13	1.52	203 (90.6%)	202 (90.2%)	0.02	(-0.17,0.17)	0.96	(0.52,1.48)
Race Caucasian	212 (89.5%)	589 (82.4%)	0.20	0.65	200 (89.3%)	200 (89.3%)	0.00	(-0.18,0.18)	1.00	(0.54,1.46)
BMI_B	31.57 (±7.24)	31.28 (±6.94)	0.04	1.09	31.52 (±7.02)	31.26 (±7.09)	0.04	(-0.19,0.19)	0.98	(0.73,1.27)
DxDur	6.5 (±6.26) /N=199/	5.1 (±6.02) /N=637/	0.23	1.08	6.2 (±5.83) /N=194/	5.8 (±6.77) /N=196/	0.06	(-0.2,0.2)	0.74	(0.6,1.4)
DrSpecialty Other Specialty	57 (24.1%)	115 (16.1%)	0.20	1.35	50 (22.3%)	52 (23.2%)	-0.02	(-0.19,0.19)	0.97	(0.75,1.25)
DrSpecialty Primary Care	37 (15.6%)	113 (15.8%)	-0.01	0.99	37 (16.5%)	33 (14.7%)	0.05	(-0.18,0.18)	1.10	(0.66,1.34)
PhysicalSymp_B	15.29 (±5.13)	13.82 (±4.54)	0.30	1.27	15.05 (±5.02)	15.34 (±4.74)	-0.06	(-0.17,0.17)	1.12	(0.72,1.28)
BPIPain_B	6.05 (±1.58)	5.46 (±1.77)	0.35	0.80	6.07 (±1.6)	6.15 (±1.85)	-0.05	(-0.16,0.16)	0.75	(0.75,1.25)
BPIInterf_B	6.7 (±1.92)	5.91 (±2.11)	0.39	0.83	6.62 (±1.91)	6.63 (±2.06)	-0.00	(-0.16,0.16)	0.86	(0.74,1.26)
FIQ_B	57.6 (±12.62)	54.1 (±13.37)	0.27	0.89	57.57 (±12.53)	58.33 (±12.73)	-0.06	(-0.17,0.17)	0.97	(0.71,1.29)
PHQ8_B	14.69 (±5.99)	12.94 (±5.84)	0.29	1.05	14.65 (±6.03)	14.78 (±5.84)	-0.02	(-0.17,0.17)	1.07	(0.79,1.21)
GAD7_B	10.93 (±5.72)	10.61 (±5.62)	0.06	1.03	10.93 (±5.71)	11.11 (±5.79)	-0.03	(-0.19,0.19)	0.97	(0.82,1.18)
CPFQ_B	27.85 (±6.43)	26.46 (±6.32)	0.22	1.03	27.92 (±6.42)	28.08 (±6.27)	-0.02	(-0.19,0.19)	1.05	(0.77,1.23)
ISIX_B	19.41 (±5.63)	17.7 (±5.5)	0.31	1.05	19.19 (±5.67)	19.27 (±5.18)	-0.01	(-0.17,0.17)	1.20	(0.7,1.3)
SDS_B	20.33 (±7.03)	18.1 (±7.38)	0.31	0.91	20.13 (±7.09)	19.98 (±6.92)	0.02	(-0.16,0.16)	1.05	(0.76,1.24)

Assess balance statistics from all METHODS

Variance ratio



Standardized Difference



Assess balance statistics from all METHODS

Propensity Method	Covariates	Method	Exact	Caliper	Ratio	Matched Treated	Matched Control	Avg Abs Standardized Differences	Max Abs Standardized Differences	% Standardized Differences < 0.1	% Standardized Differences < 0.05
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.052	0.089	100	33.333
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.042	0.099	100	73.333
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.023	0.092	100	80
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	338	0.09	0.206	60	40
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	229	229	0.091	0.232	60	40
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.072	0.167	66.667	40
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.073	0.17	60	53.333
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	317	0.058	0.223	86.667	66.667
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.053	0.118	86.667	60
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.043	0.09	100	60
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.034	0.09	100	66.667
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	342	0.078	0.219	66.667	53.333

Propensity Method	Covariates	Method	Exact	Caliper	Ratio	Matched Treated	Matched Control	Avg Abs Standardized Differences	Max Abs Standardized Differences	% Standardized Differences < 0.1
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.052	0	100
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240			
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	338			
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	229	229	0.091	0.232	60
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.072	0.167	66.667
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.073	0.17	60
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	317	0.058	0.223	86.667
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.053	0.118	86.667
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.043	0.09	100
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.034	0.09	100
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	342	0.078	0.219	66.667

Summarising the balancing scores

Show how many matches

NEW statistics
single summaries of amount of balance achieved

Propensity Method	Covariates	Method	Exact	Caliper	Ratio	Matched Treated	Matched Control	Avg Abs Standardized Differences	Max Abs Standardized Differences	% Standardized Differences < 0.1
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.052	0.089	100
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.042	0.099	100
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.023	0.092	100
Logistic Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	338	0.09	0.206	60
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	229	229	0.091	0.232	60
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Op								
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Op								
Gradient Boosted	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Var								
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Greedy		0.5	1	238	238	0.053	0.118	86.667
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal		0.5	1	240	240	0.043	0.09	100
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Optimal	Gender	0.5	1	240	240	0.034	0.09	100
Penalized Regression	Age BMI_BL BPIPain_B BPIInterf_B BPIPain_LOCF BPIInterf_LOCF Gender DrSpecialty Race Exercise	Full (Variable Ratio)		0.5		240	342	0.078	0.219	66.667

Can quickly decide: the selection of best method at a glance
Select maximum # matched
Select lowest Average/Maximum Absolute STD Difference
Select highest % STD Diff

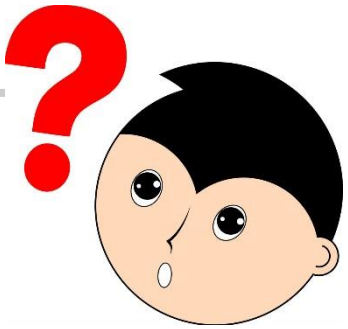
Selection of Candidate models

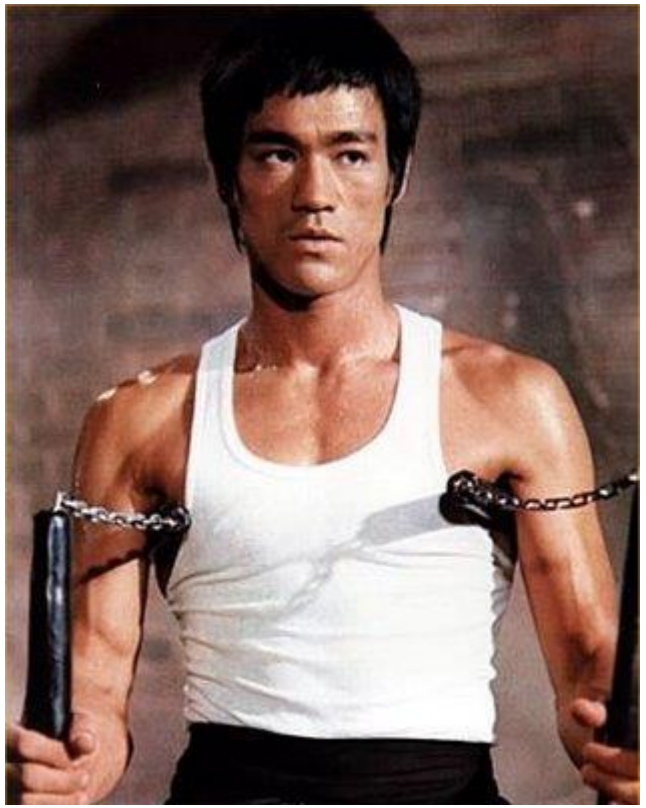
- Use the feasibility analysis:
  - Overlap
  - Balance
- To ensure the same populations are being assessed with different models
- To assess the candidate models for the model averaging

Summary

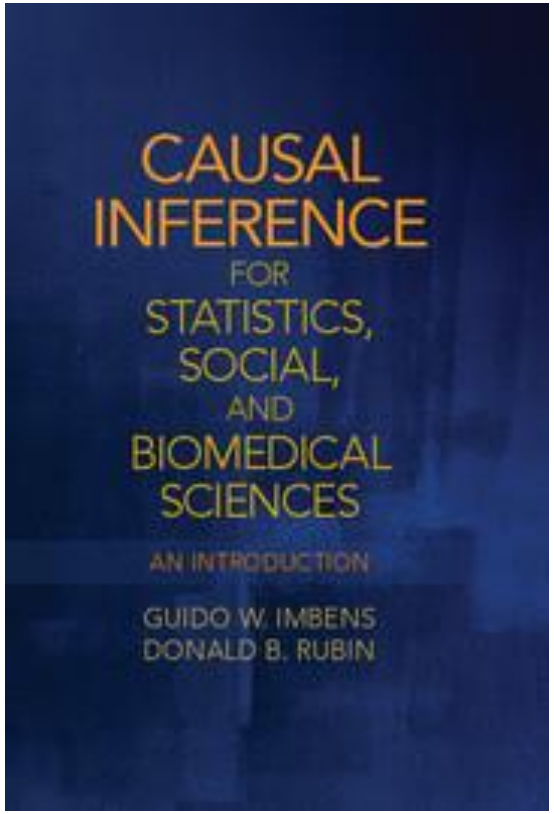
- Comparative analysis from observational data requires careful statistical adjustment for confounding
  - Process is important (pre-specification)
  - Propensity Scoring is a commonly used and useful tool to adjust for bias due to measured confounders
  - More than 1 approach is not enough
  - Advanced methods are being implemented

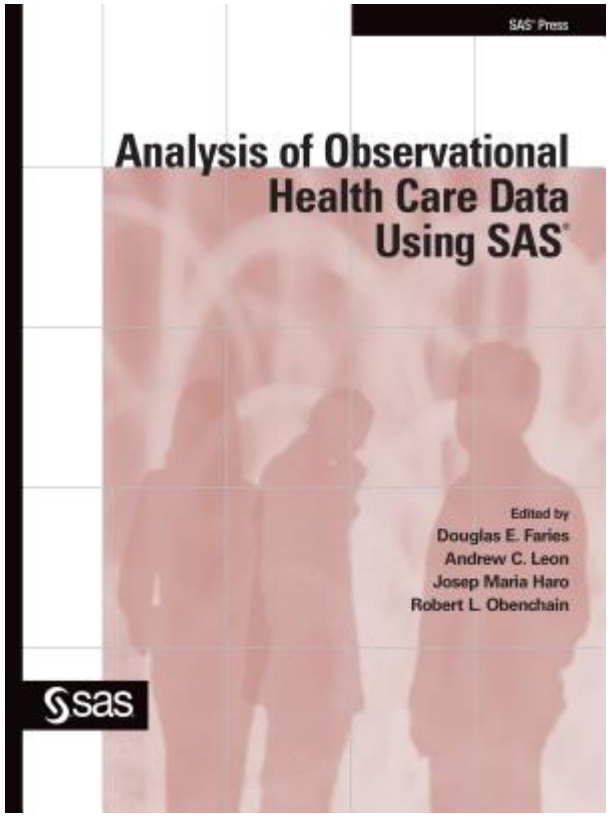
Questions

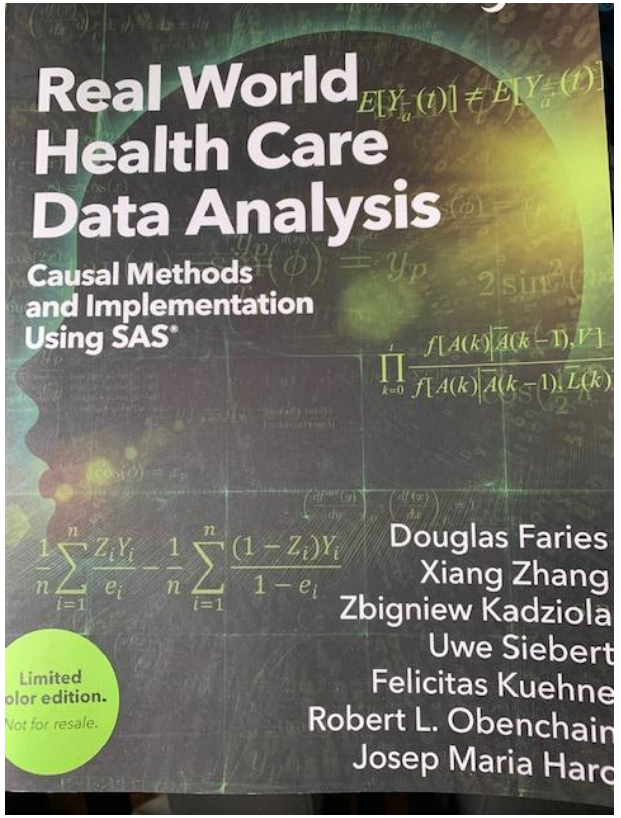




Key References







Key References

- Erik von Elm, Matthias Egger The scandal of poor epidemiological research Reporting guidelines are needed for observational epidemiology BMJ VOLUME 329 16 OCTOBER 2004
- D'Agostino RB, Kwan H: Measuring effectiveness; MedCare 1995;33;AS95-AS105
- Doshi, Jalpa A. PhD, Henry A. Glick, PhD, Daniel Polsky, PhDDoshi et al. 2006. Analyses of Cost Data in Economic Evaluations Conducted Alongside Randomized Controlled Trials. Value in health. Volume 9 Number 5 2006
- Hernán,MA. Alvaro Alonso,A. Logan,R. Grodstein,F. Michels,KB. Willett,WC. Manson,JE. James M. Robins JM. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 19(6):766-779, November 2008.
- Leigh Tooth, Robert Ware, Chris Bain, David M. Purdie, and Annette Dobson. Quality of Reporting of Observational Longitudinal Research Am J Epidemiol 2005;161:280–288
- Pocock et al (2004); Issues in the reporting of epidemiological studies. BMJ 2004;329;883
- Thelle DS, Strandhagen E: Coffee and disease: an overview. Scandinavian Journal of Nutrition 2005; 49 (2): 50-61.
- Weitzen, S. Kate L. Lapane, KL. Toledano, AY. Hume, AL. and Mor, V. Principles for modeling