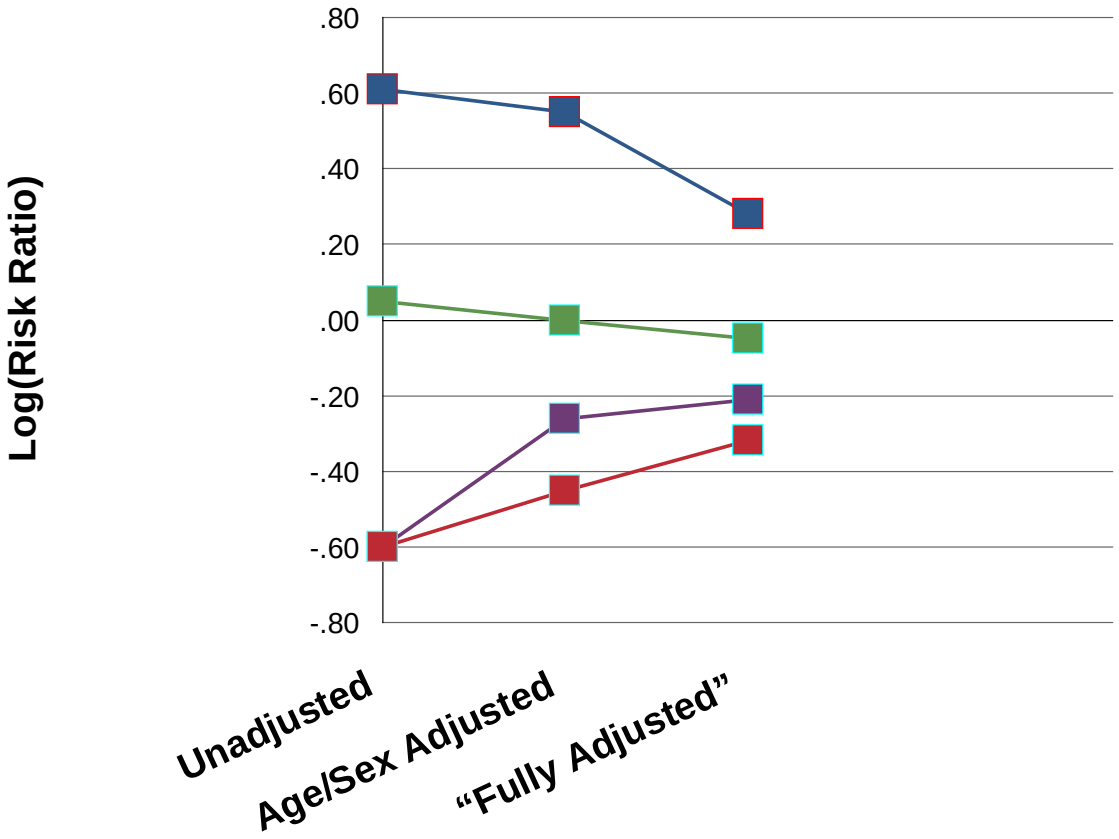
High Dimensional Propensity Scores in an Automated Setting

Jeremy A. Rassen

May 2013

Div. of Pharmacoepidemiology
Dept. of Medicine
Brigham & Women's Hospital
Harvard Medical School

Four non-randomized studies

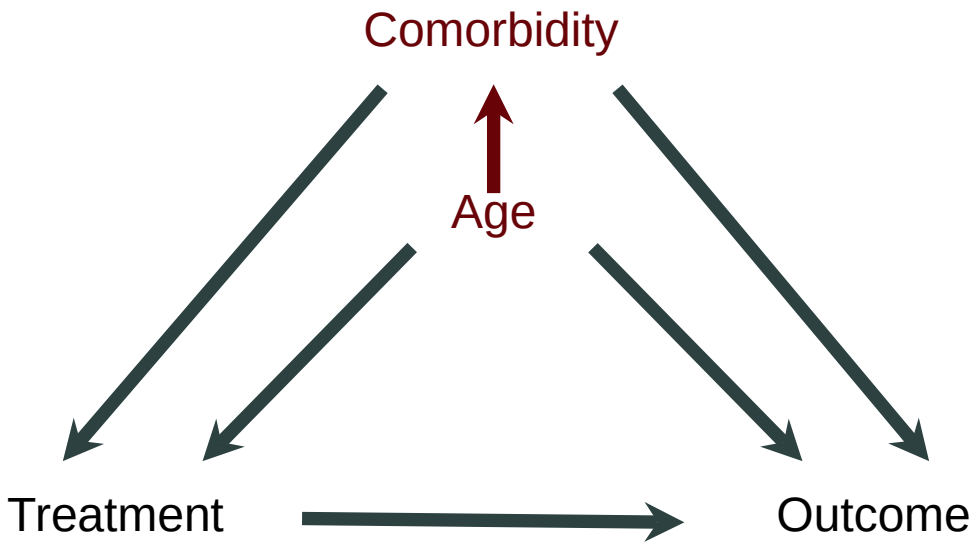


… what now?

Some approaches

- **Either remove confounders** (restriction, etc.) **or remove the effect of confounding.**
- Use approaches to deal with unmeasurable confounding
  - Instrumental variables
  - External adjustment
  - [Trials]
- Measure some previously unmeasured confounders
  - Where to get the data?

Search all the data available for confounders proxies



- Measured confounders (such as age) serve as proxies for unmeasured factors (such as general state of health).

Other examples of proxies

Observable Proxy	Implied Health State
Use of oxygen canister	Very frail
Hypertension diagnosis during hospital stay	Not too bad off
Annual checkup and colonoscopy	Careful with health; compliant patient
Incident statin use at age 70	Fairly healthy
Many drugs used, many office visits	Fairly ill

An uncomfortable thought?

- It may not matter what you're measuring as long as you can tell whether you're measuring something "important".

Automated variable identification

- Identify variables that appear to be confounders, even if they're only proxies.
- Measure as many of these as possible and adjust for them.
- A technical problem: you can't adjust for too many variables in a model (roughly 1:10 per *outcome event*)

High-Dimensional Propensity Scores

Discover Magazine, July 2012

Propensity scores and proxies

- The approach:
  - Collect as many codes as possible
  - Identify those codes that could possibly bias the exposure/outcome relationship
  - Combine variables identified *a priori* with the "best" of these codes in a propensity score.
  - Use this "high dimensional propensity score" to adjust for confounding.
- Currently implemented in a SAS macro or R program.

Where can new variables come from?

Data type

- Inpatient Diagnoses *
- Outpatient Diagnoses *
- Inpatient Procedures **
- Outpatient Procedures **
- Medication dispensings ***
- Lab test results
- Unstructured text notes

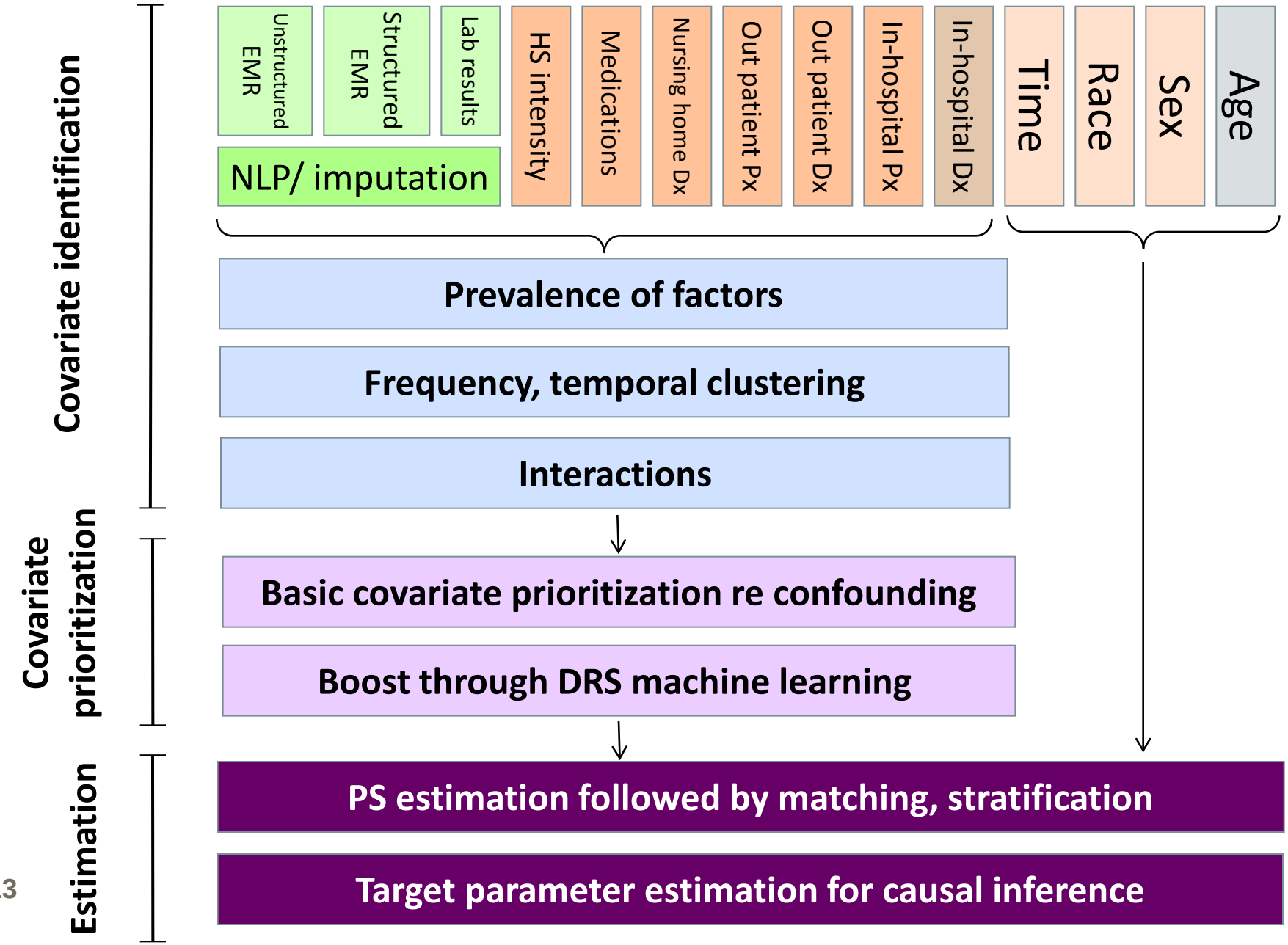
Frequency/ Intensity

- Once
- Sporadic
- Frequent

Temporality

- Proximal to exposure
- Evenly distributed
- Distal to exposure start

High-dimensional propensity score



hd-PS algorithm (1/7)

- **STEP 1a. Specify Data Sources**
  Define *p* data dimensions; use data stream of 180 days up to the initiation of study exposure. Collect all codes from each of the *p* dimensions.

- **STEP 1b.**
  Include basic demographic information (age, sex) and investigator-defined covariates (history of diabetes, use of statins, ...)

hd-PS algorithm (2/7)

- **STEP 2. Identify empirical candidate covariates**
  Within each data dimension sort by prevalence of codes. Identify the *n* most prevalent codes and how the codes should be treated.

- ***Example:***
  - *n*= 200
  - Use 3 digit ICD-9 (vs. 4 or 5 digit)
  - Use 5 digit CPT
  - Use generic drug name (vs. drug class)

hd-PS algorithm (3/7)

- **STEP 3. Assess code recurrence and create indicator variables for each patient**

  For each identified code, address frequency by creating 3 variables:

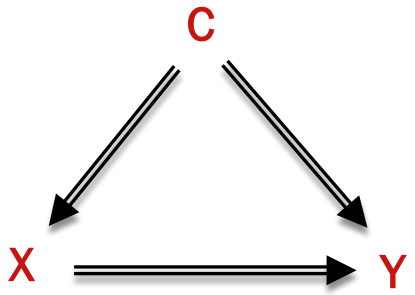
  - CovX_once = 1 if that code appeared at least once within 180 days
  - CovX_sporadic = 1 if code appeared at least more than the median
  - CovX_frequent = 1 if code appeared at least more than the 75th percentile.

hd-PS algorithm (4/7)

- **STEP 4. Prioritize covariates**
  Calculate for each covariate the possible amount of confounding the covariate could adjust for.

  Use the Bross (1966) formula:

$$BIAS_{mult} = \frac{P_{C1}(RR_{CY} - 1) + 1}{P_{C0}(RR_{CY} - 1) + 1}$$



hd-PS algorithm (5/7)

- **STEP 5. Select covariates**
  Add the demographic covariates and investigator-defined covariates from step 1. Then, select top *k* empirical covariates from step 4 as ranked by potential bias.

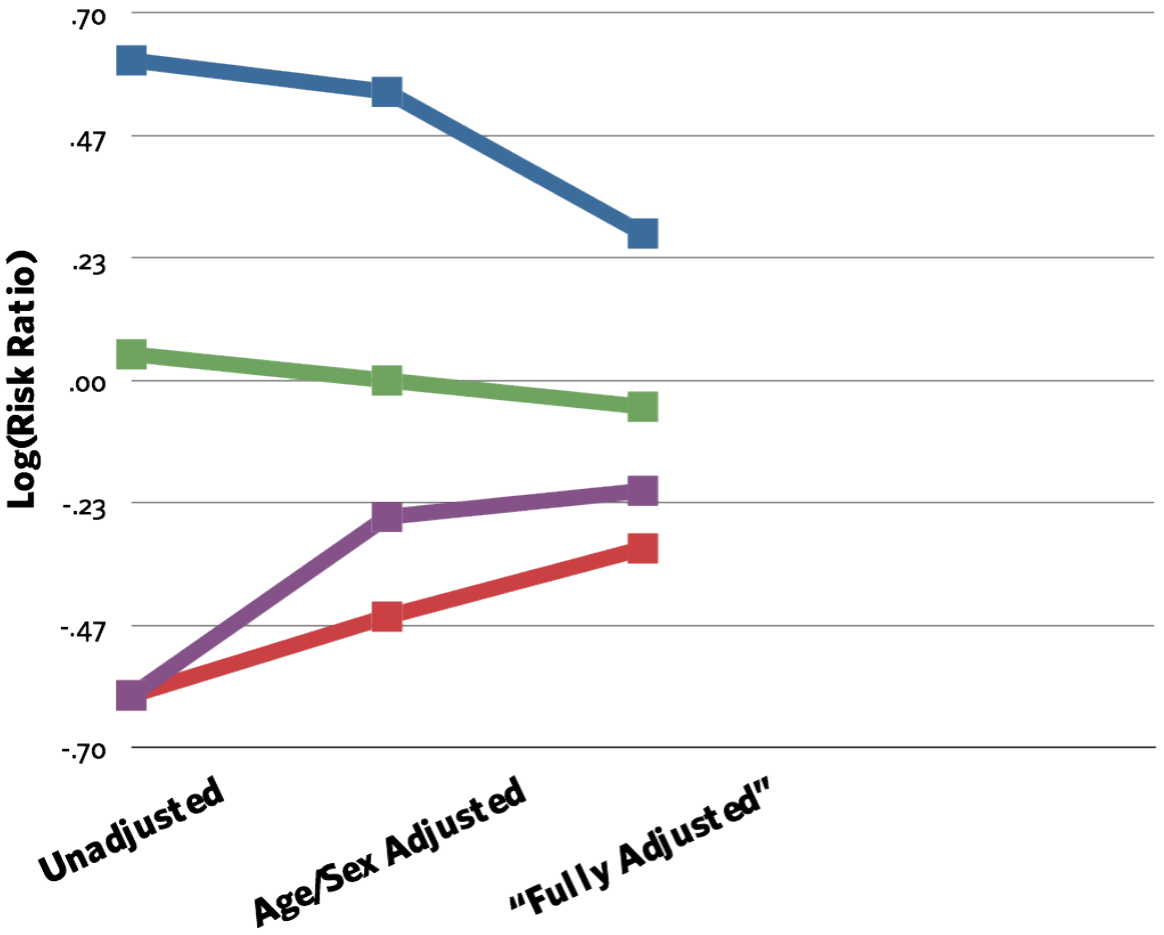
- **Example:**
  - 4 demographic covariates
  - 24 investigator-defined covariates
  - *k*=500 hd-PS-selected covariates

hd-PS algorithm (6 + 7/7)

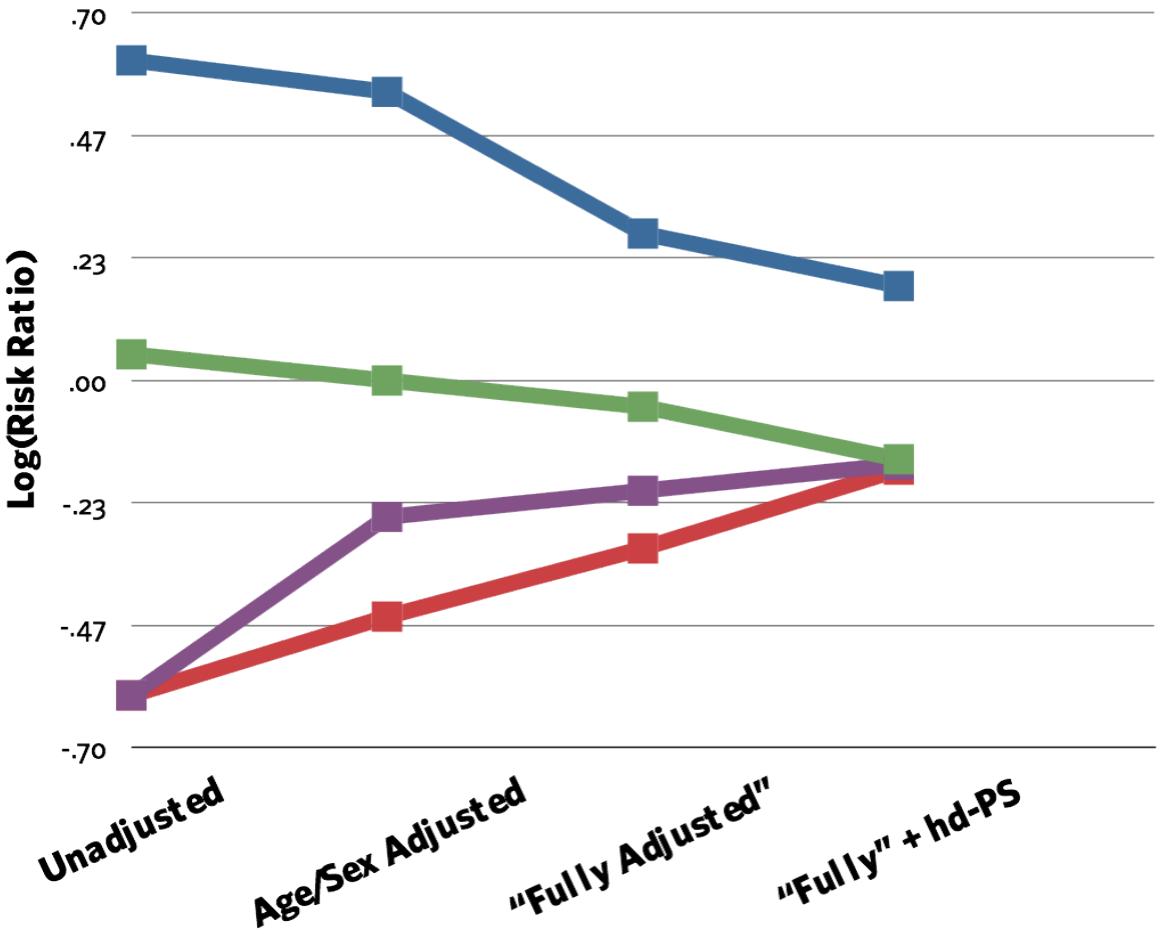
- **STEP 6. Estimate propensity score (the hd-PS)**
  Estimate propensity score using multivariate logistic regression, including all investigator-defined covariates and the k hd-PS-selected covariates.

- **STEP 7. Use the hd-PS**
  Use the hd-PS in an outcome model as any PS would be used (matching, deciling, trimming, etc.)

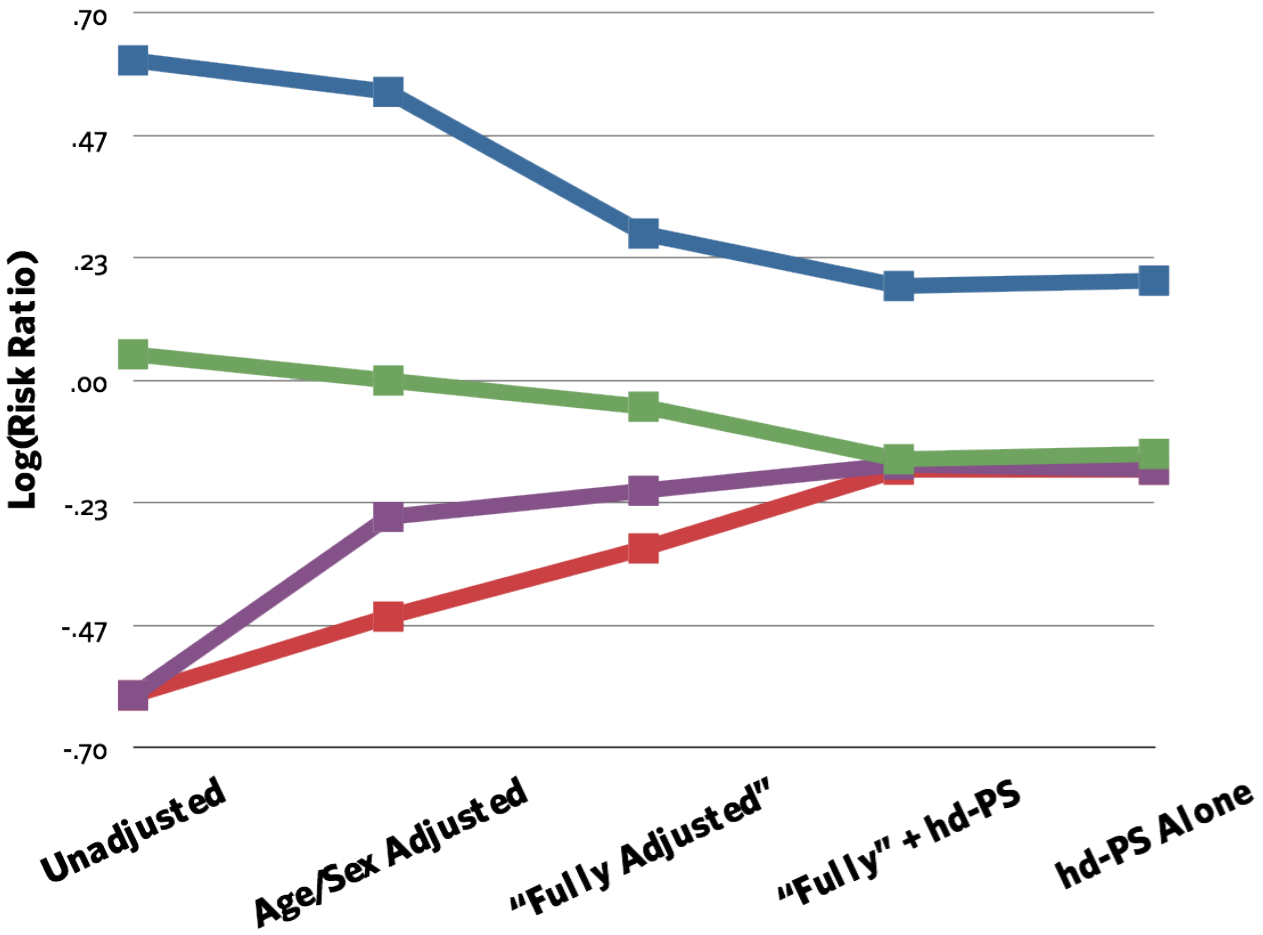
Recall the four studies



Recall the four studies



Recall the four studies



hdPS empirical performance

Data sources

Insurance claims data:

U.S. Medicare

U.S. commercial

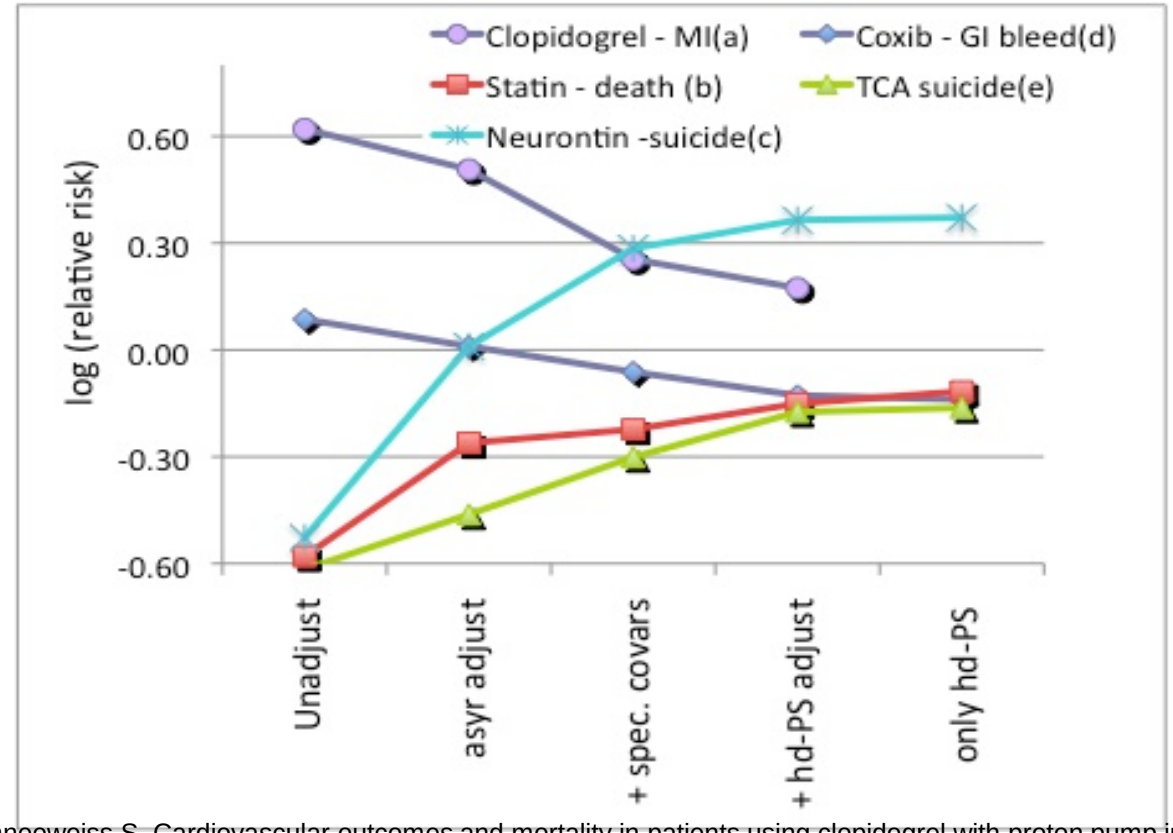
Canada

Germany

Elect. health records:

United Kingdom

Regenstrief



(a) Rassen JA, Choudhry N, Avorn J, Schneeweiss S. Cardiovascular outcomes and mortality in patients using clopidogrel with proton pump inhibitors after percutaneous coronary intervention. Circulation 2009;120:2322-9.

(b) Schneeweiss S, Rassen JR, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology 2009;20:512–22.

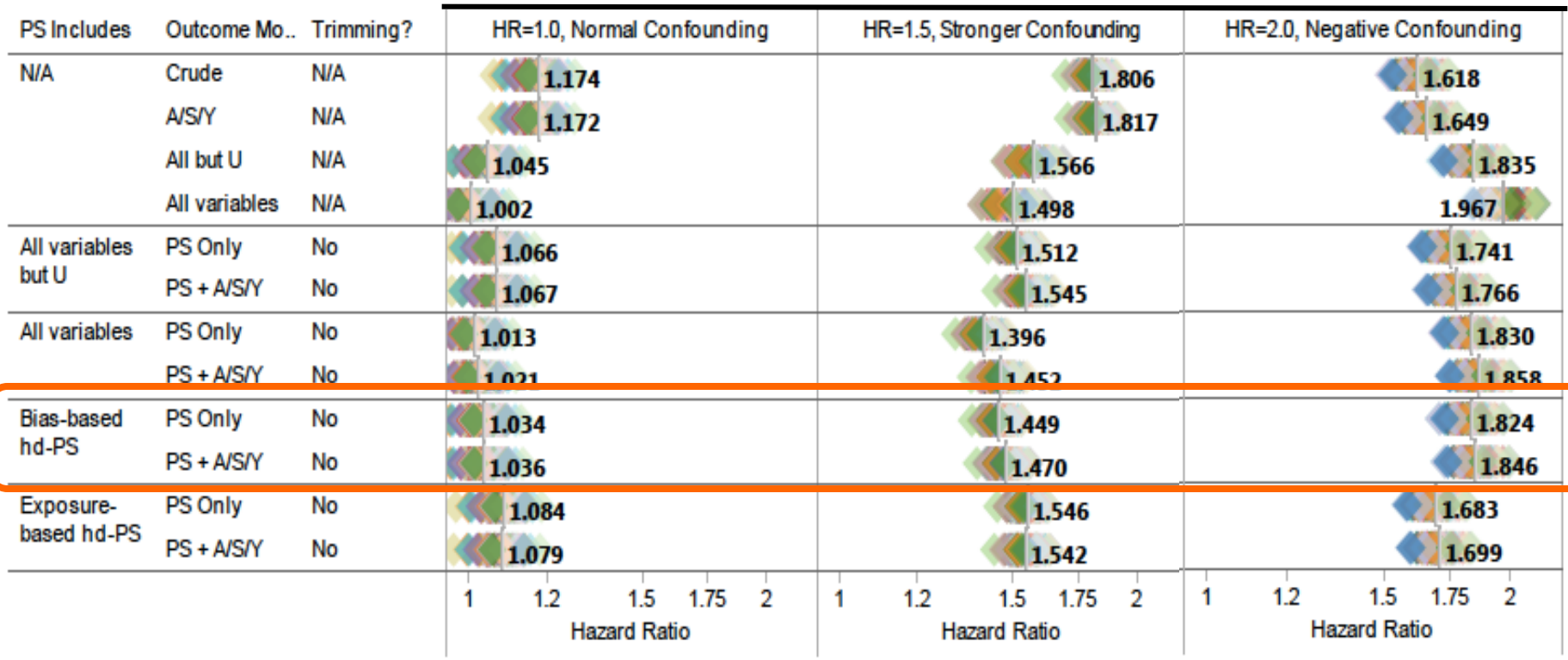
(c) Patorno E, Bohn RL, Wahl PM, Avorn J, Patrick AR, Liu J, Schneeweiss S. Anticonvulsant medications and the risk of suicide, attempted suicide, or violent death. JAMA 2010;303:1401-9

(d) Same as (b)

(e) Schneeweiss S, Patrick AR, Solomon DH, Metha J, Dormuth C, Miller M, Lee J, Wang PS. The comparative safety of antidepressant agents in children regarding suicidal acts. Pediatrics 2010;125: 876–88

hdPS simulation performance

Simulated true effect estimate* (HR= 1.0, 1.5, 2.0)



* Plasmode simulations inject a defined causal effect of E on Y|C in a given healthcare database preserving the underlying data structure and information content.

Getting the software

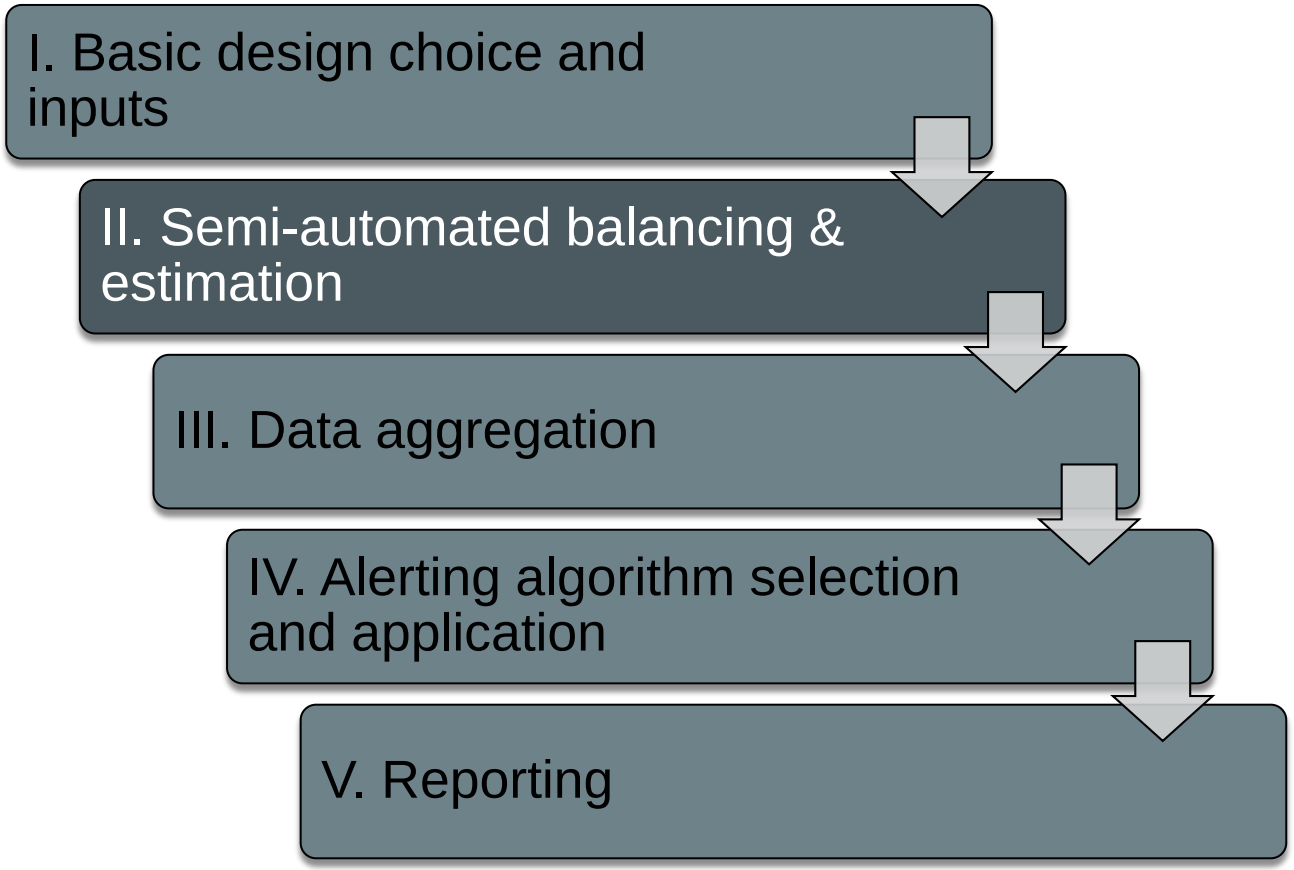
- All materials available for download at: http://www.drugepi.org

- Open source (Mozilla 2.0 license)

Some more examples

Exposure (referent)	Outcome	Relative risk estimate (95%CI)			
		Unadjusted	Adjusted by basic variables	Basic variables plus other pre-specified variables	Basic variables plus hd-PS
Coxibs (ns-NSAIDS) [7]	GI bleed within 180 days	1.09 (0.91–1.30)	1.01 (0.84–1.21)	0.94 (0.78–1.12)	0.88 (0.73–1.06)
Coxibs (ns-NSAIDS) [41]	GI bleed	1.21 (0.91–1.61)	–	0.99 (0.74–1.33)	0.67 (0.45–0.97)
Coxibs (ns-NSAIDS) [13]	Upper GI bleed	1.50 (0.98–2.28)	0.84 (0.54–1.31)	0.81 (0.52, 1.27)	0.78 (0.49, 1.22)
Statins (glaucoma drugs) [7]	All-cause mortality within 180 days	0.56 (0.51–0.62)	0.77 (.069–0.85)	0.80 (0.70–0.90)	0.86 (0.76–0.98)
TCAs (SSRIs), <18 y.o. [9]	Suicide within 1 year	0.59 (0.28–1.27)	0.66 (0.31–1.42)	0.71 (0.33–1.52)	0.92 (0.43–2.00)
TCA (SSRIs), 18+ y.o. [10]	Suicide within 1 year	0.97 (0.77–1.21)	1.04 (0.83–1.31)	1.04 (0.82–1.31)	1.14 (0.88–1.47)
Clopidogrel + PPI (clopidogrel alone) [42]	MI or CV death	1.74 (1.44–2.10)	1.62 (1.34–1.96)	1.32 (1.08–1.61)	1.22 (0.99–1.51)
Neurontin (Topamax) [11]	Suicide or attempted suicide	0.95 (0.76–1.19)	1.48 (1.17–1.87)	1.42 (1.11–1.80)	1.99 (1.45–2.73)
Conventional APMs (atypical APMs) [12]	Death	1.37 (1.11–1.69)	1.47 (1.13–1.90)	1.47 (1.14–1.91)	1.52 (1.14–2.02)
Benzodiazepines (atypical APMs) [12]	Death	1.37 (1.14–1.64)	1.52 (1.25–1.85)	1.28 (1.04–1.58)	1.20 (0.96–1.50)

Using hd-PS in a sequential monitoring environment



Example *Vioxx*

- Drug of interest: *rofecoxib*
- Comparators: *non-selective NSAIDs*
- Other exposure parameters:
  - *New users (180-day washout)*
  - *"As-treated" with 30-day maximum gap*
- Outcome of interest: *AMI (inpatient ICD-0 410.x1)*
- Pre-defined covariates:
  - *18 ICD-9 based conditions*
  - *9 drug classes*
- hdPS options: *defaults + health service utilization variables*
- Databases: *PACE and PAAD, 5/20/1999 to 12/31/2004*

Diagnostics *Outputs*

- Goals of diagnostics:
  - To identify errors and areas of missing data
  - To assess level of balance achieved
  - Evaluate and compare patient characteristics, numbers of events, rates across data partners
  - To decide whether to include data in ongoing monitoring
- Table 1: *demographics, predefined covariates, M-distance*
- Figure 1: *PS distribution, c-statistics*
- Additional figures: *summary and detailed balance figures*

* All of these can be produced for unmatched, matched, or stratified cohorts

Diagnostics *Vioxx Table 1*

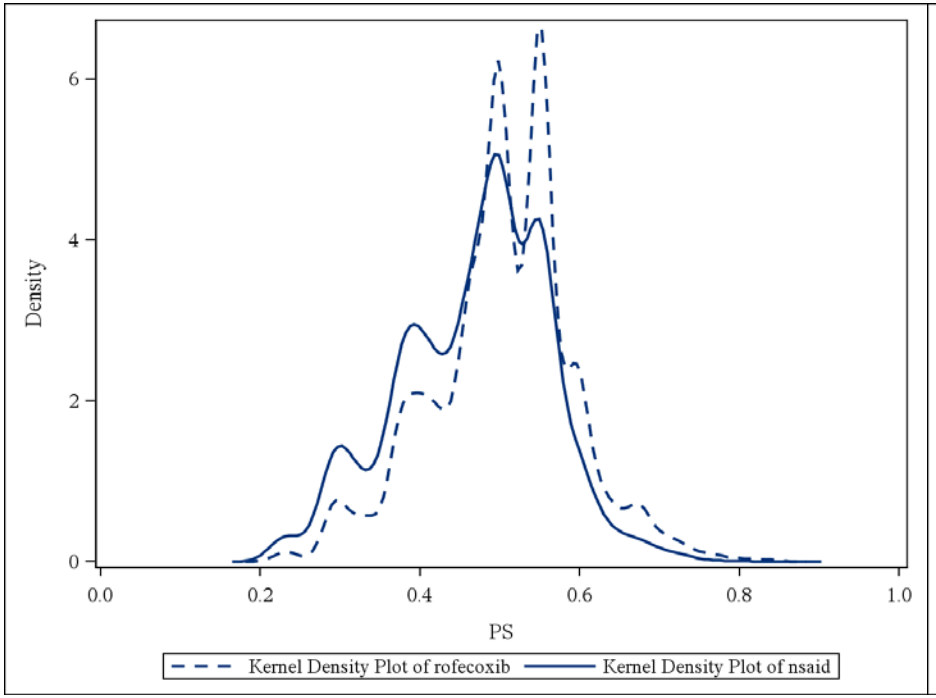
Table 1. Cohort of New Initiators of Rofecoxib and Non-Selective NSAID (Unmatched)

	Primary Analysis		Covariate Balance	
			Absolute	Standardized
	N (%)	N (%)	Difference	Difference
Characteristic	rofecoxib	nsaid		
Number of patients	9409 (100.0 %)	9977 (100.0 %)		
Number of Events While on Therapy	39 (0.4 %)	15 (0.2 %)		
Person time at risk	59.9 (33.3)	46.4 (32.5)		
Patient Characteristics				
Age	76.3 (10.7)	73.1 (12.2)	3.2	3.2
60-70	1305 (13.9 %)	1679 (16.8 %)	-2.9	-0.082
70-80	3631 (38.6 %)	3883 (38.9 %)	-0.3	-0.007
80-90	3179 (33.8 %)	2619 (26.3 %)	7.5	0.164
90-100	580 (6.2 %)	395 (4.0 %)	2.2	0.101
Gender (F)	7764 (82.5 %)	7374 (73.9 %)	8.6	0.208
Recorded use of:				
Ace Inhibitors	1224 (13.0 %)	1351 (13.5 %)	-0.5	-0.016
ARB	567 (6.0 %)	535 (5.4 %)	0.6	0.029
Anticoagulants	548 (5.8 %)	328 (3.3 %)	2.5	0.122
Mahalanobis Distance	0.131			

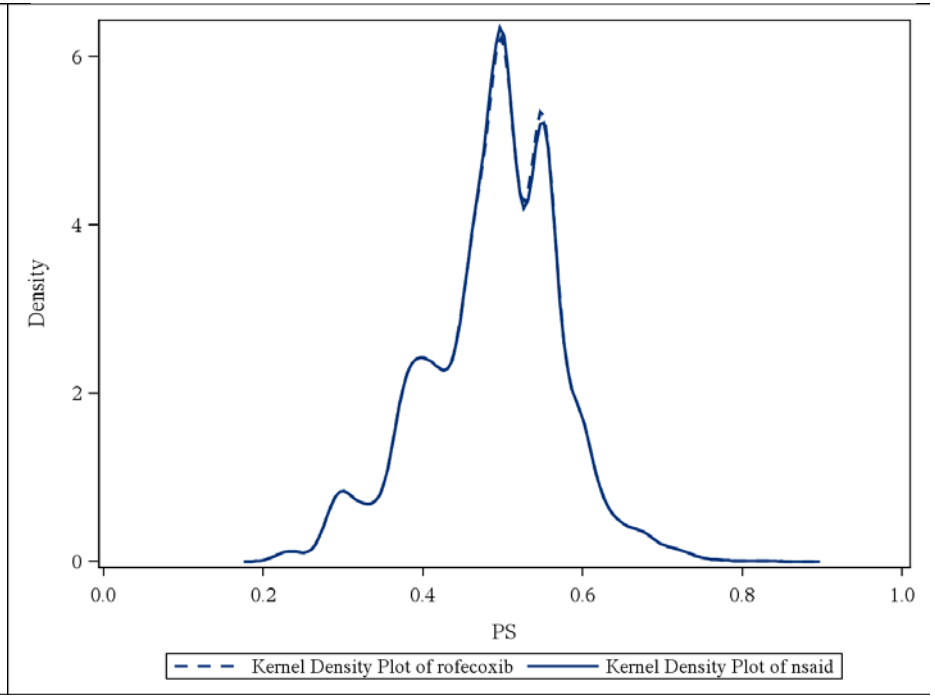
Additional predefined covariates

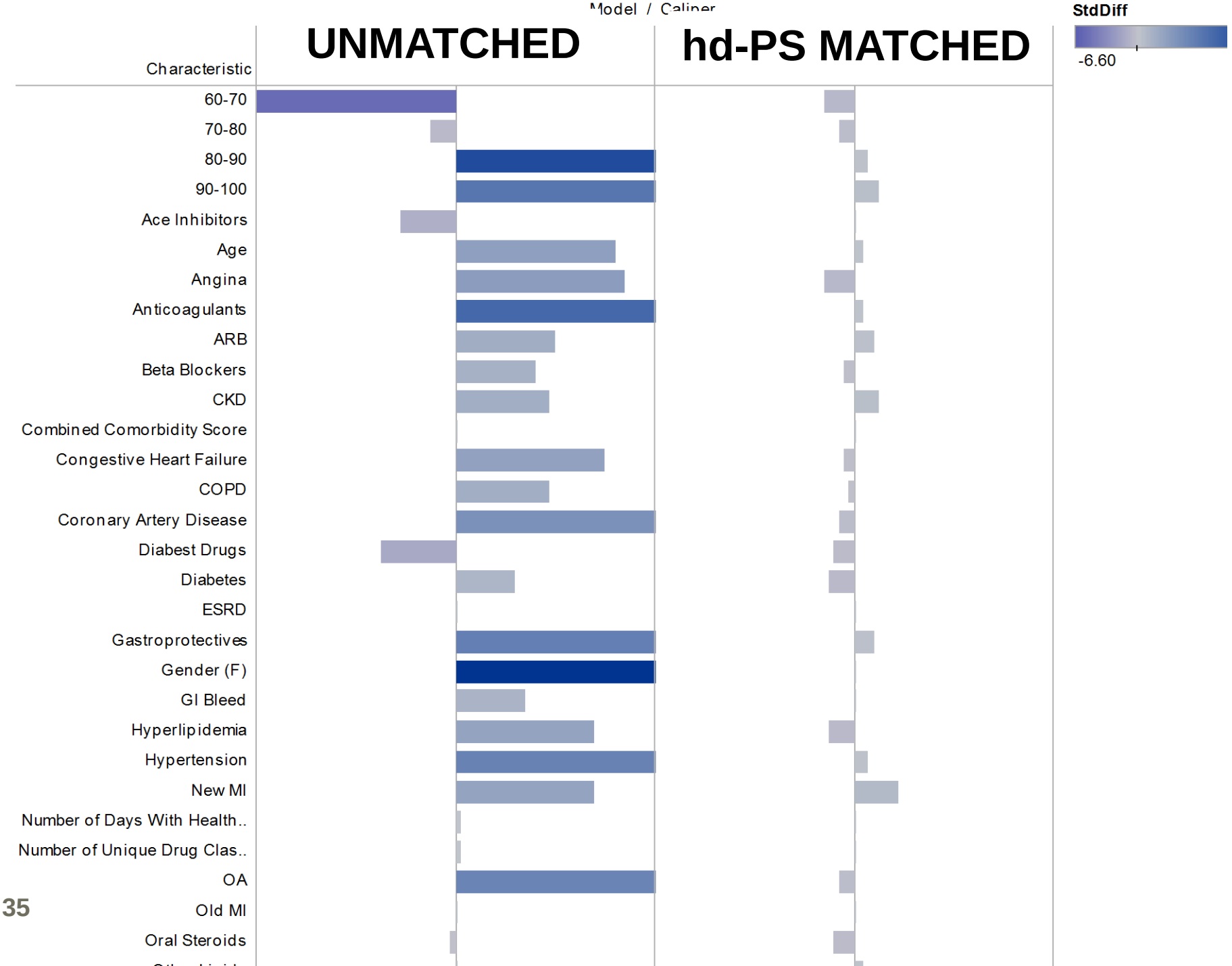
Diagnostics: match quality

UNMATCHED



hd-PS MATCHED




Model / Caliper
UNMATCHED
hd-PS MATCHED
StdDiff
-6.60
Characteristic
60-70
70-80
80-90
90-100
Ace Inhibitors
Age
Angina
Anticoagulants
ARB
Beta Blockers
CKD
Combined Comorbidity Score
Congestive Heart Failure
COPD
Coronary Artery Disease
Diabest Drugs
Diabetes
ESRD
Gastroprotectives
Gender (F)
GI Bleed
Hyperlipidemia
Hypertension
New MI
Number of Days With Health..
Number of Unique Drug Clas..
OA
Old MI
Oral Steroids

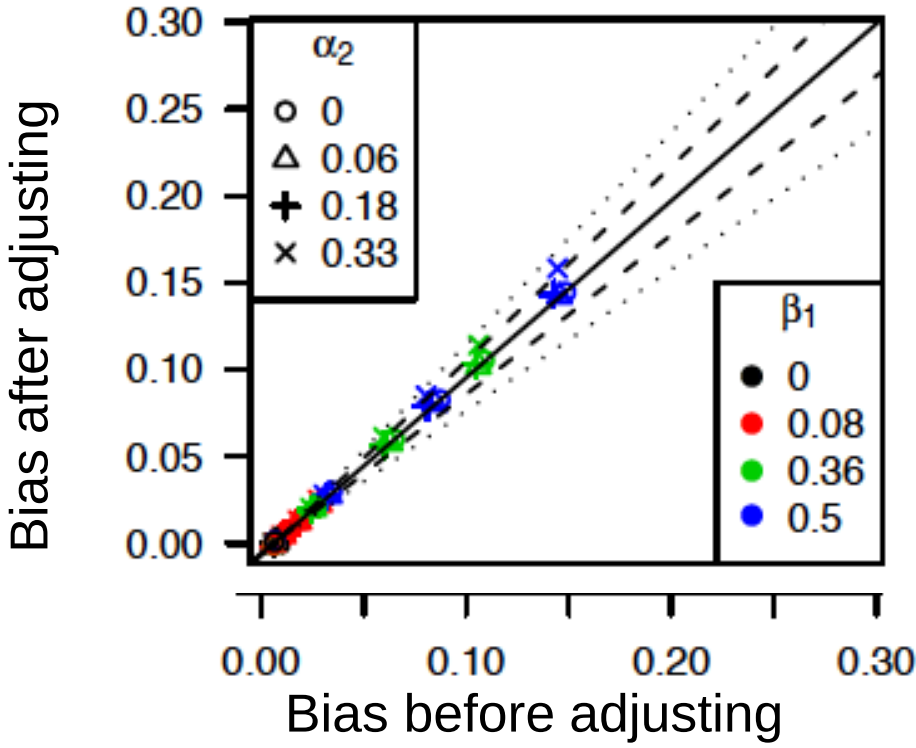
Issues arising in automatic PS variable selection

- Variable selection is often based on the strength of association with treatment.
- Will tend to select IVs (instrumental variables) or near-IVs, variables that strongly predict treatment but have a weak association with outcome.
- Balancing IVs and near-IVs may amplify any residual confounding bias[1,2] in some cases.

[1] Battacharya & Vogt, *NBER Technical Working Paper* no. 343, 2007.
[2] Pearl, *Proceedings of the UAI 2010*.

Issues arising in automatic PS variable selection

- Simulation studies showed that increases in bias due to adjusting for IVs were generally small (< 10%) except in extreme examples[1].



When in doubt whether a covariate is an IV, include the covariate. Unmeasured confounding bias is worse than Z-bias.

[1] Myers, Rassen, Gagne, et al., *AJE*, 2011;174(11):1213-1222.

Issues with PSs with new drugs

- In early experience with a new drug, a prescriber's use is likely to evolve
  - Need to re-estimate propensity scores after brief time periods to model changes in therapeutic choice
- Emergence of new evidence on efficacy and safety in specific patient groups can also induce changes over time in propensity scores.

Issues with PSs with new drugs (cont.)

- Even for drugs that go on to be blockbusters, initial time periods after approval will have few users, limiting ability to select variables for and fit high-dimensional propensity scores.
- Distinction of true risk factors from variables that influence treatment choice but not outcome (instruments) is challenging with infrequently used drugs

Schneeweiss, S., Gagne, J. J., Glynn, R. J., Ruhl, M. & Rassen, J. A. Assessing the Comparative Effectiveness of Newly Marketed Medications: Methodological Challenges and Implications for Drug Development. *Clinical Pharmacology & Therapeutics* 1–14 (2009).doi:10.1038/clpt.2011.235

Thank you.