

High-dimensional propensity scores for comparative effectiveness research

Joshua J. Gagne

Division of Pharmacoepidemiology
Department of Medicine
Brigham and Women's Hospital
Harvard Medical School

Lewin Summit
June 15, 2010

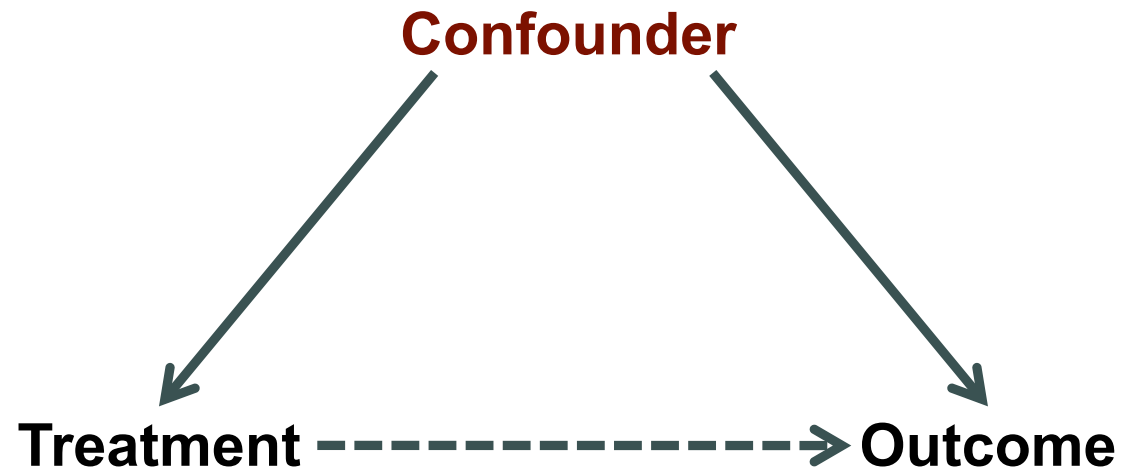


Disclosure

- ▶ Employee of Brigham and Women's Hospital and have been partially funded by the Pharmacoepidemiology program at the Harvard School of Public Health

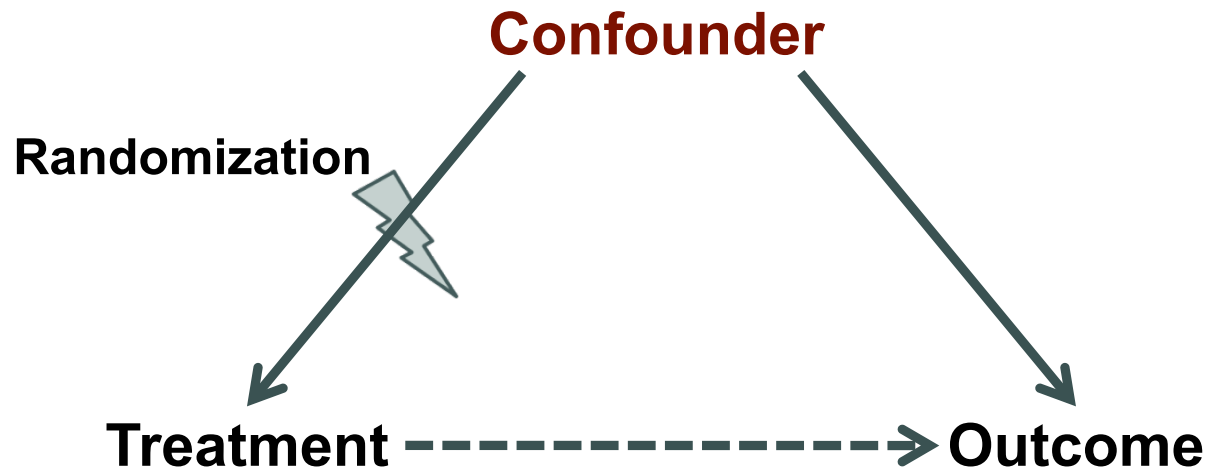
Confounding

- ▶ Patient factors that are associated with treatment choice and are risk factors for the outcome are confounders

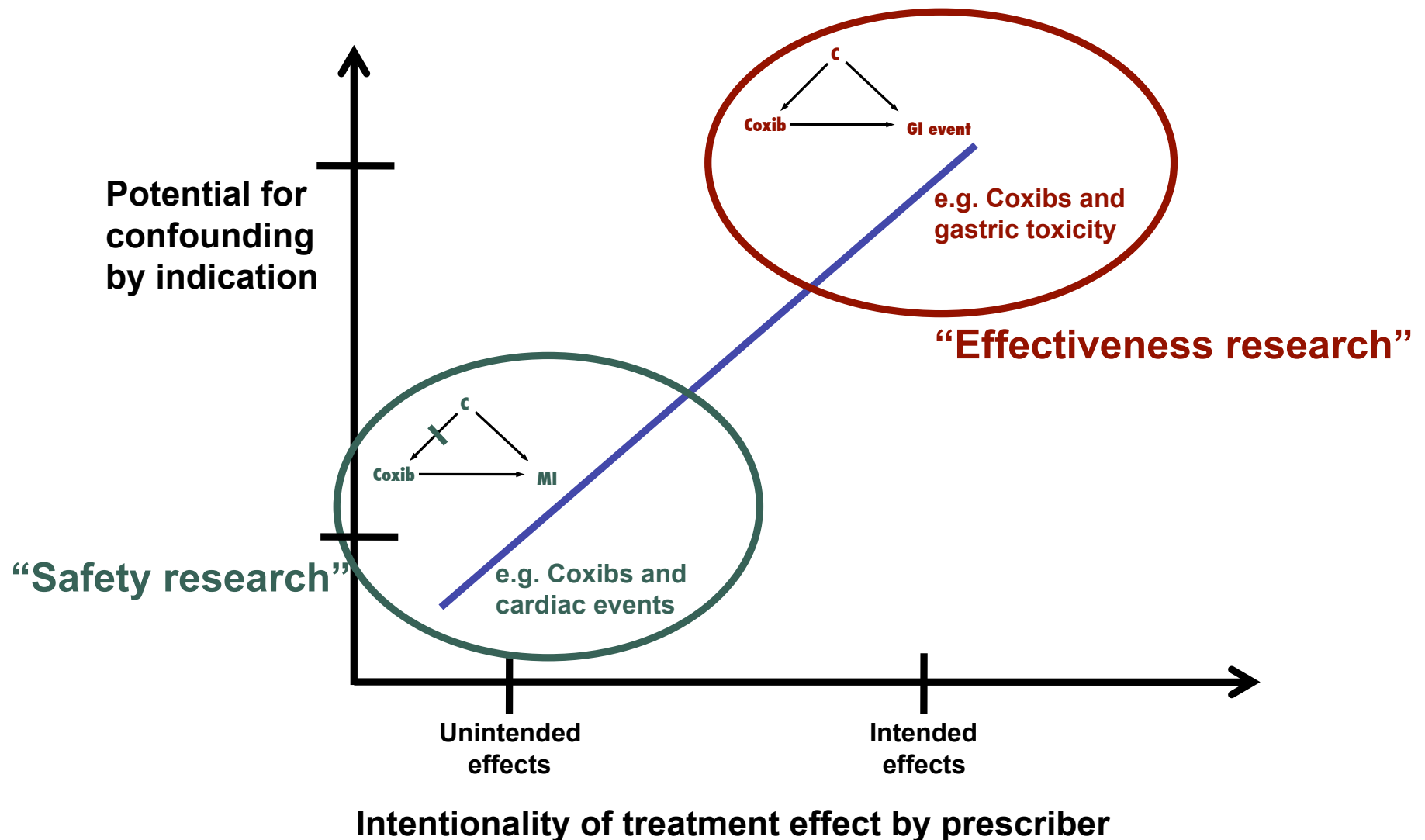


Confounding

- ▶ Patient factors that are associated with treatment choice and are risk factors for the outcome are confounders



Treatment intention and confounding

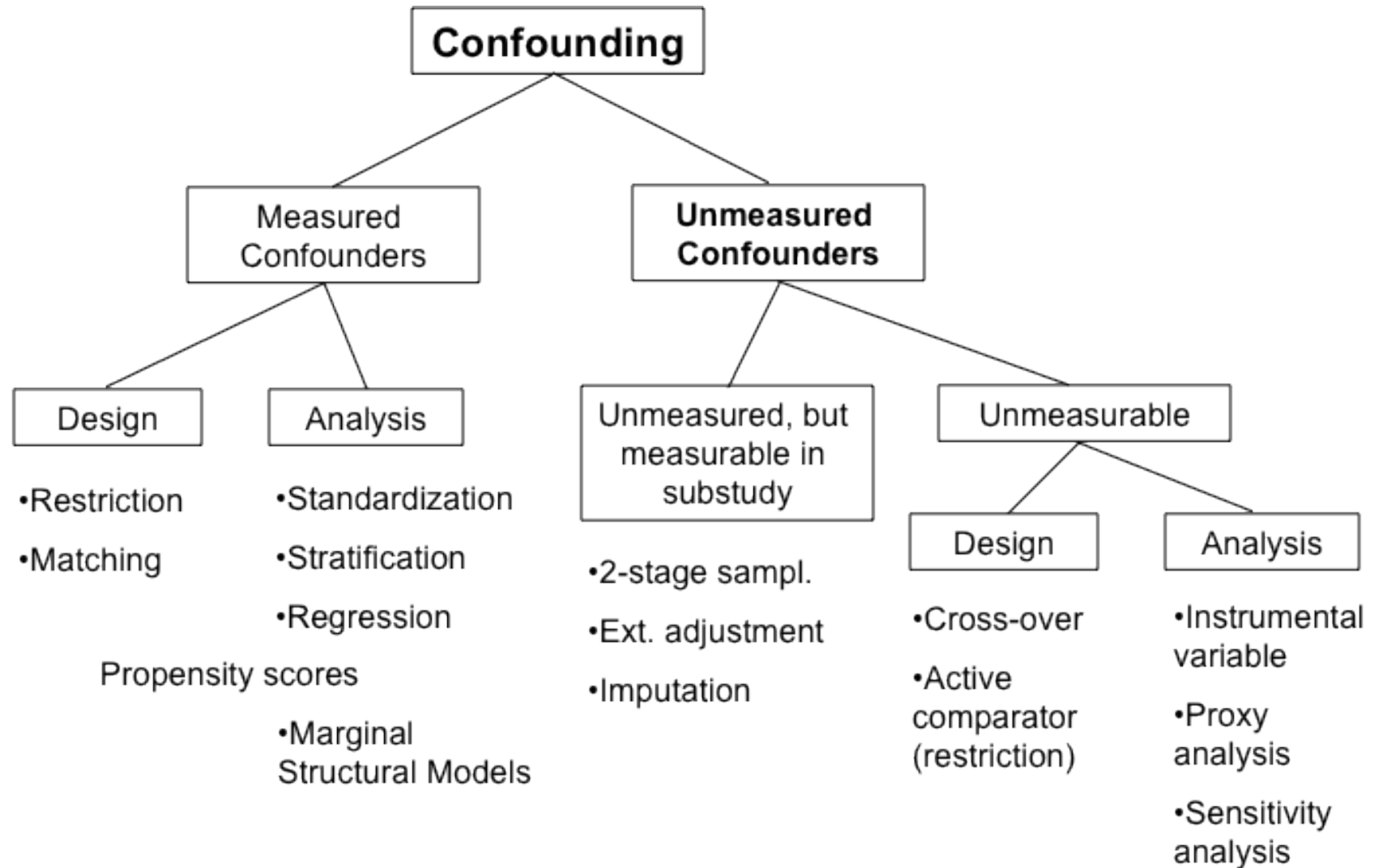


In other words

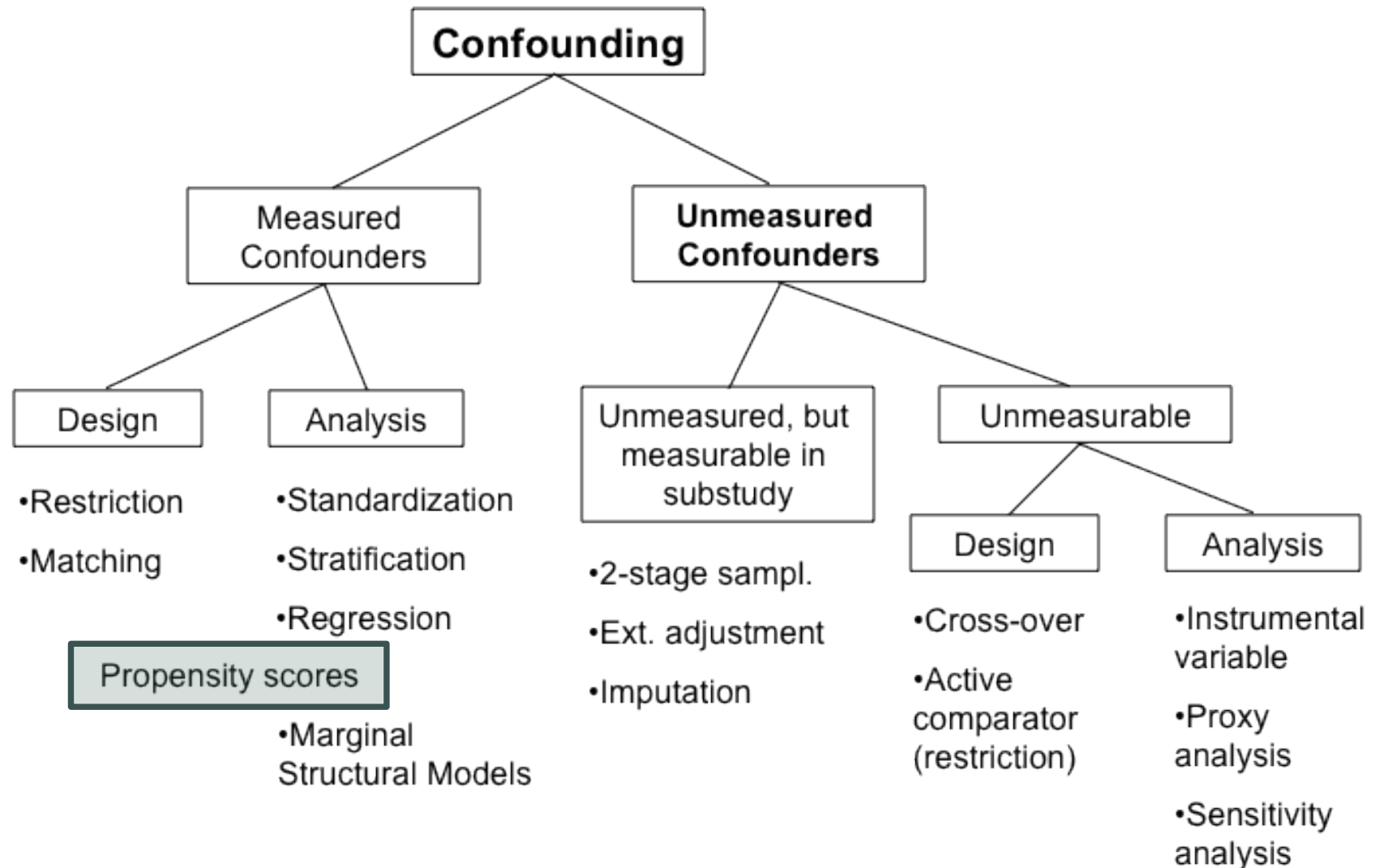
“There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don’t know. But there are also unknown unknowns. These are things we do not know we don’t know.”

- Donald Rumsfeld (February 12, 2002)

Confounding control in observational data



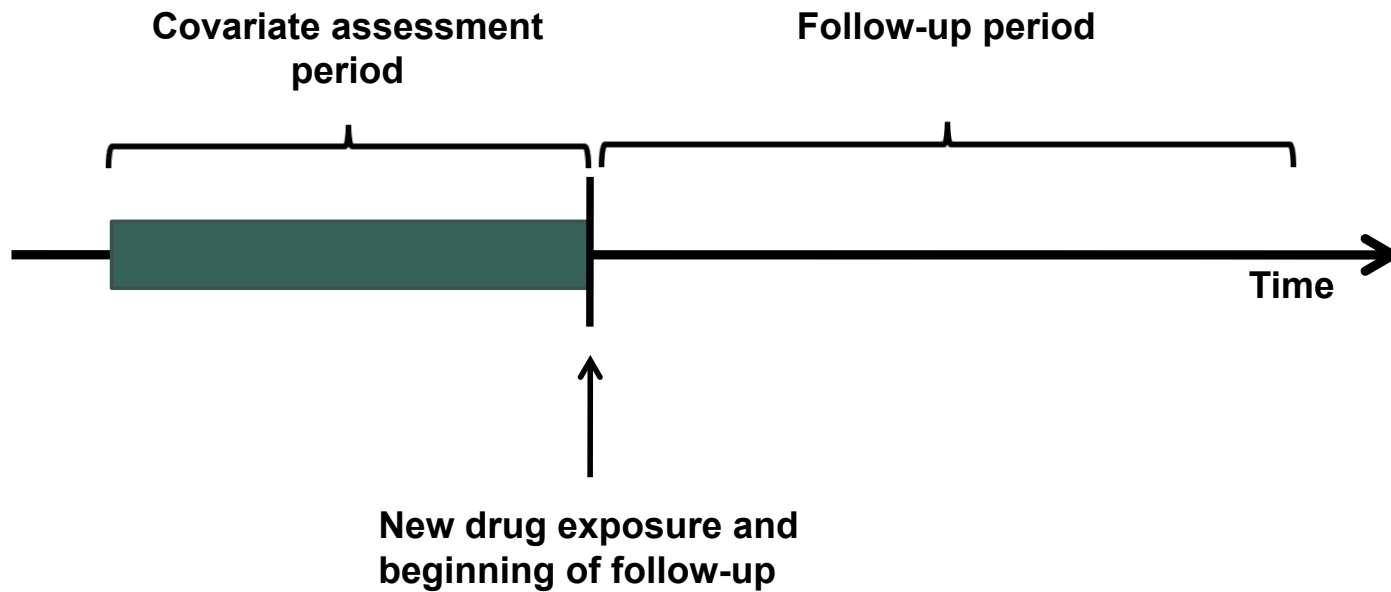
Confounding control in observational data



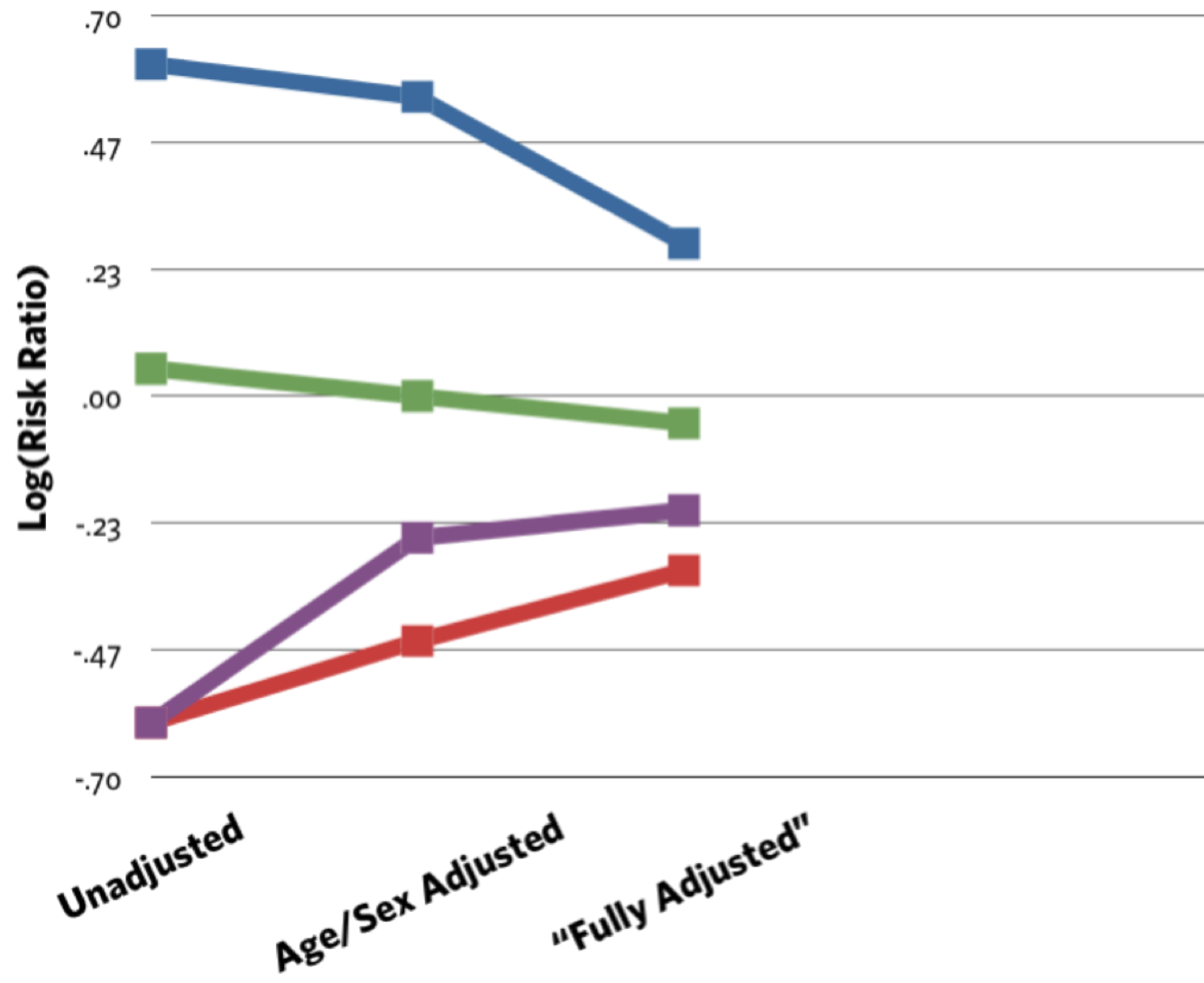
Propensity score analysis

- ▶ Goal: to identify patients with the same likelihood of receiving treatment among those treated and untreated
- ▶ Two-stage process:
 - ▶ Estimate the propensity for treatment as a function of observed covariates
 - ▶ Mimic the treatment decision process
 - ▶ Predicted value is each patient's "propensity score"
 - ▶ Use the estimated PS to adjust the treatment model:
 - ▶ Quantile stratification
 - ▶ Match by PS
 - ▶ Outcome model adjustment

Propensity scores in cohort studies



Incremental adjustment



Richness of administrative data

- ▲ Constant flow of data with little delay and at low cost
- ▲ Millions of patients with defined person–time denominator
- ▲ Data reflect routine care
- ▲ Generalizable to large population segments
- ▲ HIPAA compliance protects patient privacy

Claims data

- **Member ID**
- Plan
- Gender
- Age
- Dates of Eligibility

- **Member ID**
- Prescribing physician
- Drug dispensed (NDC)
- Quantity and date dispensed
- Drug strength
- Days supply
- Dollar amounts

- **Member ID**
- Physician or Facility identifier
- Procedures (CPT-4, revenue codes, ICD-9)
- Diagnosis (ICD-9-CM, DRG)
- Admission and discharge dates
- Date and place of service
- Dollar amounts

Administrative data

Pharmacy claims data

Physician and facility claims data

Supplemental data

- **Member ID**
- Lab Test Name
- Result

Lab test results data

- **Member ID**
- Income
- Net Worth
- Education
- Race & Ethnicity
- Life Stage
- Life Style Indicators

Consumer elements

- **Member ID**
- Subspecialty notes
- Endoscopy reports
- Histology reports
- Radiology reports
- Free text notes

Electronic medical records

Computerized, linked, longitudinal dataset

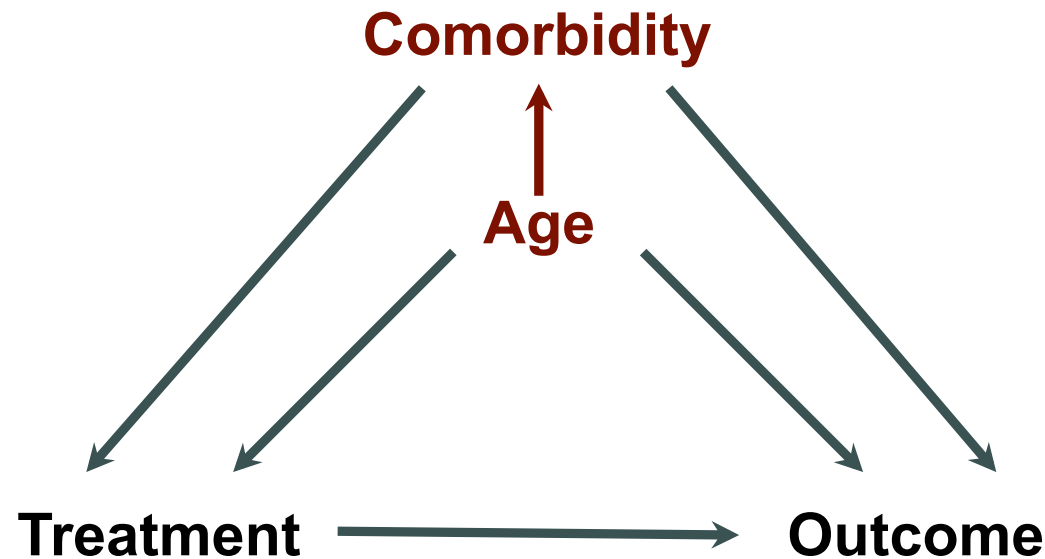
An abundance of codes

----- ID=***** dob=**/**/1948 sex=M eligdt=1/2000 indexdt=6/2001 -----

Service Date	Site of Service	Prov Type	Code	Drug or Procedure Description	* Code	Diagnosis Description
10/01/00	OFFICE	Family Practice	90658	INFLUENZA VIRUS VACC/SPLIT	V048	VACC FOR INFLUEN
10/01/00	Rx	Pharmacy		CIPROFLOXACIN 500MG TABLETS	10	
11/05/00	OFFICE	Family Practice	17110	DESTRUCT OF FLAT WARTS, UP	0781	VIRAL WARTS
11/07/00	Rx	Pharmacy		CIPROFLOXACIN 500MG TABLETS	10	
01/15/01	Rx	Pharmacy		CIPROFLOXACIN 500MG TABLETS	10	
06/25/01	OFFICE	Emerg Clinic	99070	SPECIAL SUPPLIES	* 84509	SPRAIN OF ANKLE
					E927	ACC OVEREXERTION
06/30/01	OFFICE	Orthopedist	99204	OV,NEW PT.,DETAILED H&P,LOW	* 72767	RUPT ACHILL TEND
06/30/01	OFFICE	Internist/Gener	99202	OV,NEW PT.,EXPD.PROB-FOCSD	* 84509	SPRAIN OF ANKLE
	OUTPT HP	Anesthesiologis	01472	REPAIR OF RUPTURED ACHILLES	* 84509	SPRAIN OF ANKLE
		Hospital	27650	REPAIR ACHILLES TENDON	* 84509	SPRAIN OF ANKLE
			85018	BLOOD COUNT; HEMOGLOBIN	* 84509	SPRAIN OF ANKLE
		Orthopedist	27650	REPAIR ACHILLES TENDON	* 84509	SPRAIN OF ANKLE
06/30/01	OFFICE	Orthopedist	29405	APPLY SHORT LEG CAST	* 72767	RUPT ACHILL TEND
07/30/01	OFFICE	Orthopedist	29405	APPLY SHORT LEG CAST	* 72767	RUPT ACHILL TEND
08/13/01	OFFICE	Orthopedist	L2116	AFO TIBIAL FRACTURE RIGID	* 72767	RUPT ACHILL TEND

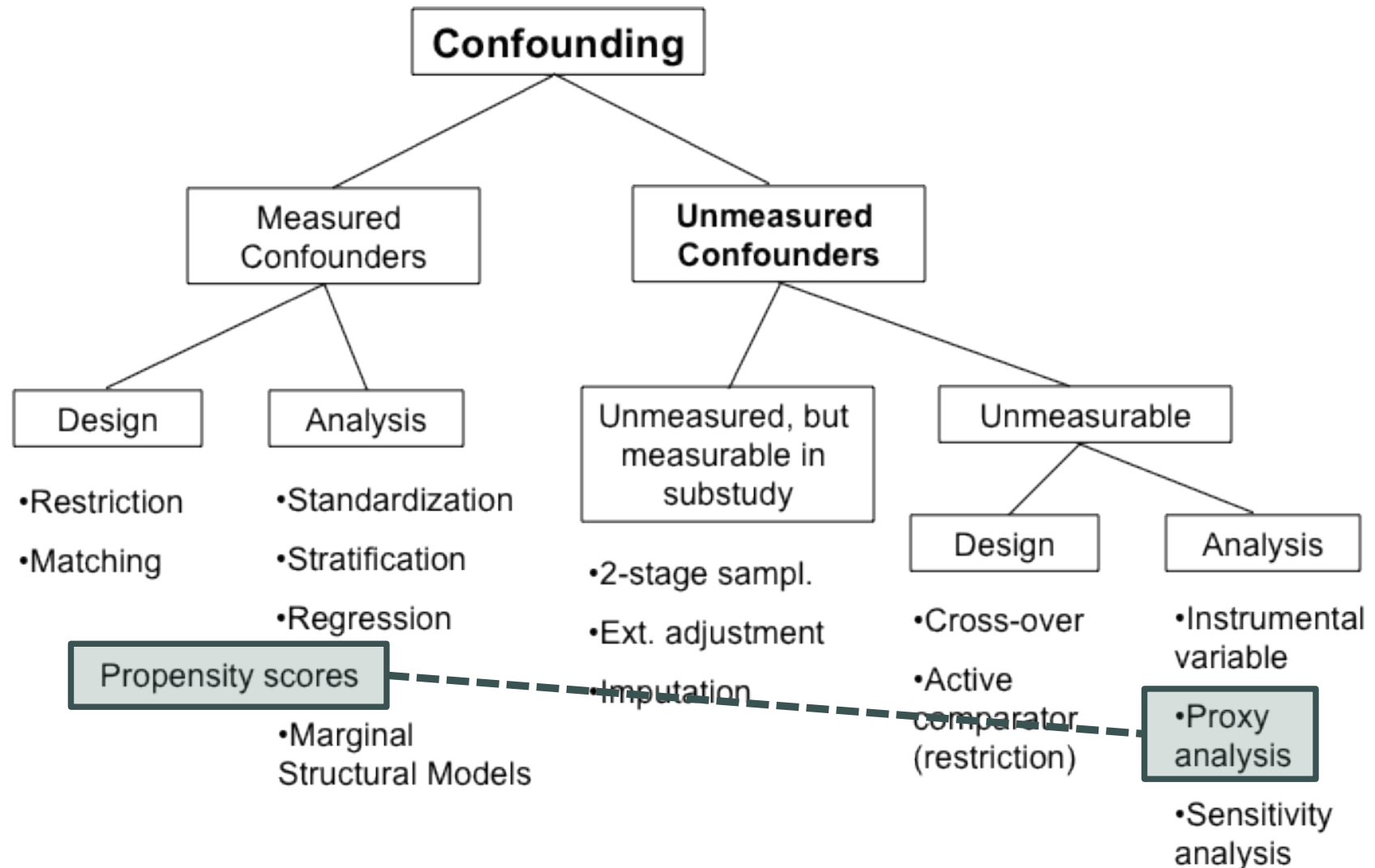
Proxies for confounders

Basic idea: measured confounders (such as age) serve as proxies for unmeasured factors (such as general state of health)

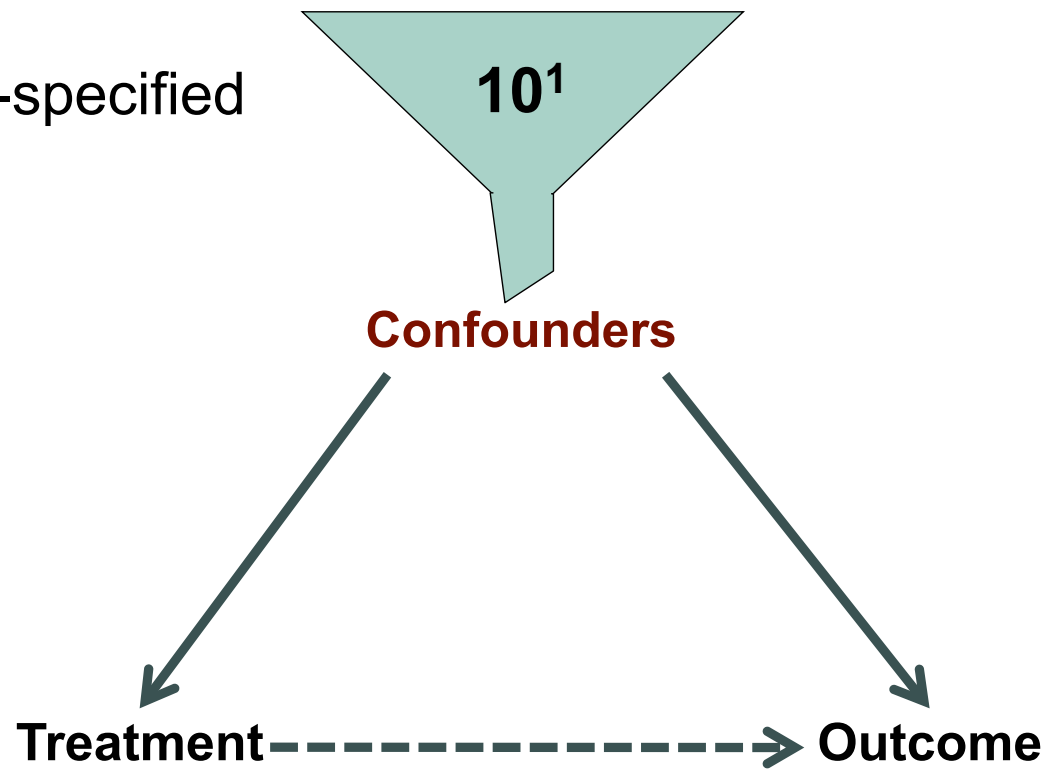


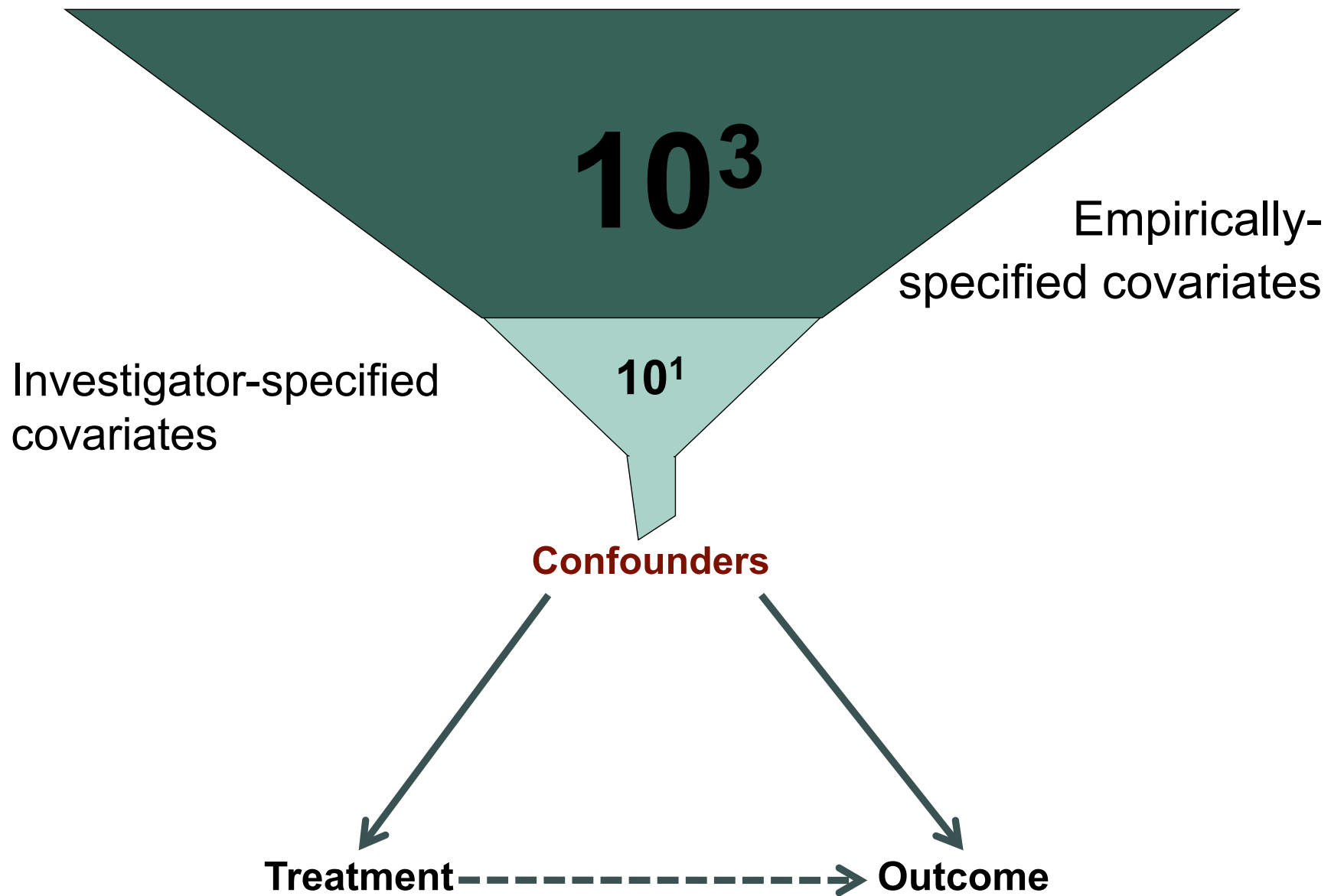
Finding proxies for previously unmeasured confounders should lead to better confounding adjustment

Confounding control in observational data



Investigator-specified
covariates





Propensity scores and proxies

- ▶ Collect as many (diagnostic, procedural, drug) codes as possible
- ▶ Identify codes that could possibly bias the treatment/outcome relation
- ▶ 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

1,000,000-s of
data items and
patterns

Mining codes,
texts, patterns
with new and
existing tools

10,000-s of
potential
confounding
factors

Prioritization
according to
potential for
confounding

1,000-s of
confounders

Statistical ana-
lysis using e.g.
propensity
score methods
or shrinkage
estimators

The hd-PS SAS macro.

The hd-PS SAS macro can be downloaded at www.drugapi.org ... links ... downloads.

EXAMPLE CODE

```
%include "/path/to/macro/directory/hdps.mcr";
```

```
Title1 'High-dimensional propensity score adjustment';
```

```
Title2 '(study description)';
```

```
%RunHighDimPropScore (  
    var_patient_id      = id,  
    var_exposure         = exposed,  
    var_outcome          = outcome,  
    vars_demographic    = age sex race,  
    vars_force_categorical = year,  
    top_n               = 200,  
    k                   = 500,  
    trim_mode           = BOTH,  
    percent_trim        = 5,  
    input_cohort        = master_file,  
    input_dim1          = drug_claims      generic_name,  
    input_dim2          = outpatient_diagnoses icd9_dx,  
    input_dim3          = inpatient_diagnoses icd9_dx,  
    input_dim4          = inpatient_procedures icd9_proc,  
    input_dim5          = outpatient_procedures cpt,  
    output_scored_cohort = scored_cohort,  
    output_detailed     = detailed_cohort,  
    results_estimates   = estimates,  
    results_diagnostic  = variable_info  
);
```

hd-PS algorithm

Flow chart for basic high-dimensional propensity score algorithm.

1. Specify data sources

Define p data dimensions; use data stream of 180 days up to the initiation of study exposure. This includes diagnoses on the day of initiation but no drugs or procedures on the day of initiation. Exclude selected codes from covariate adjustment

Base case: $p = 8$

1	2	3	4	5	6	7	8
MC Part A*	MC Part A	MC Part A	MC Part A	MC Part A	Part B	Part B	Drugs
Hosp Dx	Hosp proc	Amb Dx	Amb proc	Nurse Home	Dx	Proc	generic
(ICD-9)	(ICD-9)	(ICD-9)	(ICD-9)	(ICD-9)	(ICD-9)	(CPT-4)	entities

* MC = Medicare, Hosp = hospital, Amb = ambulatory, Dx = diagnosis, proc = procedure, Rx = prescription dispensings

Investigator specified covariates

Demographics

Age, sex, race, year

Predefined

Hx, Dx, Rx, Procs

2. Identify empirical candidate covariates

Within each data dimension sort by prevalence of codes. Identify the n most prevalent codes.

Base case: $n = 200$; Granularity = 3 digit ICD-9, 5 digit CPT, generic drug name.

3. Assess recurrence

For each identified code create three covariates:

CovX_once = 1 if that code appeared at least once within 180 days

CovX_sporadic = 1 if code appeared at least twice

CovX_frequent = 1 if code appeared at least three times.

4. Prioritize covariates

Within each data dimension calculate for each covariate the possible amount of confounding it could adjust in a multiplicative model given a binary exposure and outcome after adjusting for demographic covariates:

$\text{Bias}_{\text{mult}} = \frac{P_{C1}(RR_{CD} - 1) + 1}{P_{C0}(RR_{CD} - 1) + 1}$ if $RR_{CD} \geq 1$, $\frac{P_{C1}(\frac{1}{RR_{CD}} - 1) + 1}{P_{C0}(\frac{1}{RR_{CD}} - 1) + 1}$ otherwise. Sort in descending order.

5. Select covariates

Add d demographic covariates from step 1 and l predefined covariates in the top positions. Select top k

in a multiplicative model given a binary exposure and outcome after adjusting for demographic covariates:

$$\text{Bias}_{\text{mult}} = \frac{P_{C1}(RR_{CD} - 1) + 1}{P_{C0}(RR_{CD} - 1) + 1} \text{ if } RR_{CD} \geq 1, \frac{P_{C1}(\frac{1}{RR_{CD}} - 1) + 1}{P_{C0}(\frac{1}{RR_{CD}} - 1) + 1} \text{ otherwise. Sort in descending order.}$$

5. Select covariates

Add d demographic covariates from step 1 and l predefined covariates in the top positions. Select top k empirical covariates from step 4. Optional, include multiplicative 2-way interactions for d demographic and l predefined covariates with the top 20 empirical covariates.

Base case: $d = 4$ (age, sex, race, year); $l = 14$; $k = 500$

6. Estimate exposure propensity score

Estimate propensity score using multivariate logistic regression, including all $d + l + k$ covariates. Truncate 5% of patients on either end of PS distribution and form deciles.

7. Estimate outcome model

Estimate exposure-outcome association adjusted for propensity score deciles as well as PS weighted.

Empirical example:

Coxibs vs. nsNSAIDs and risk of GI complications in 180 days

Table 1: Characteristics of 49,653 initiators of selective COX-2 inhibitors or non-selective (ns) NSAIDs as defined during 6 months prior to first medication use.

	Initiators of Cox-2 selective NSAIDs		Initiators of nsNSAIDs		OR*	95% CI
	N	%	N	%		
N	32,042		17,611			
Age 75 years or older	24,079	75%	11,496	65%	1.61	1.545-1.674
Sex, % female	27,528	86%	14,293	81%	1.42	1.348-1.487
Race: white	30,583	95%	15,808	90%	2.39	2.23-2.57
black	1,133	4%	1,580	9%	0.37	0.34-0.403
other	326	1%	223	1%	0.80	0.68-0.95
Charlson comorbidity score ≥ 1	24,343	76%	12,521	71%	1.29	1.233-1.340
Use of >4 distinct drugs in prior year	24,120	75%	11,852	67%	1.48	1.421-1.541
>4 physician visits in prior year	22,919	72%	11,363	65%	1.38	1.328-1.437
Hospitalized in prior year	9,804	31%	4,591	26%	1.25	1.200-1.303
Nursing home resident	2,671	8%	996	6%	1.52	1.407-1.635
Prior use of gastroprotective drugs	8,785	27%	3,600	20%	1.47	1.407-1.536
Prior use of warfarin	4,252	13%	1,153	7%	2.18	2.041-2.337
Prior use of oral steroids	2,800	9%	1,373	8%	1.13	1.059-1.211
History of OA	15,549	49%	5,898	33%	1.87	1.802-1.945
History of RA	1,602	5%	476	3%	1.90	1.707-2.102
History of peptic ulcer disease	1,189	4%	426	2%	1.55	1.389-1.739
History of gastrointestinal hemorrhage	551	2%	196	1%	1.55	1.319-1.831
History of hypertension	23,332	76%	12,363	70%	1.14	1.092-1.184
History of congestive heart failure	9,727	30%	4,328	25%	1.34	1.283-1.395
History of coronary artery disease	5,266	16%	2,603	15%	1.13	1.078-1.193

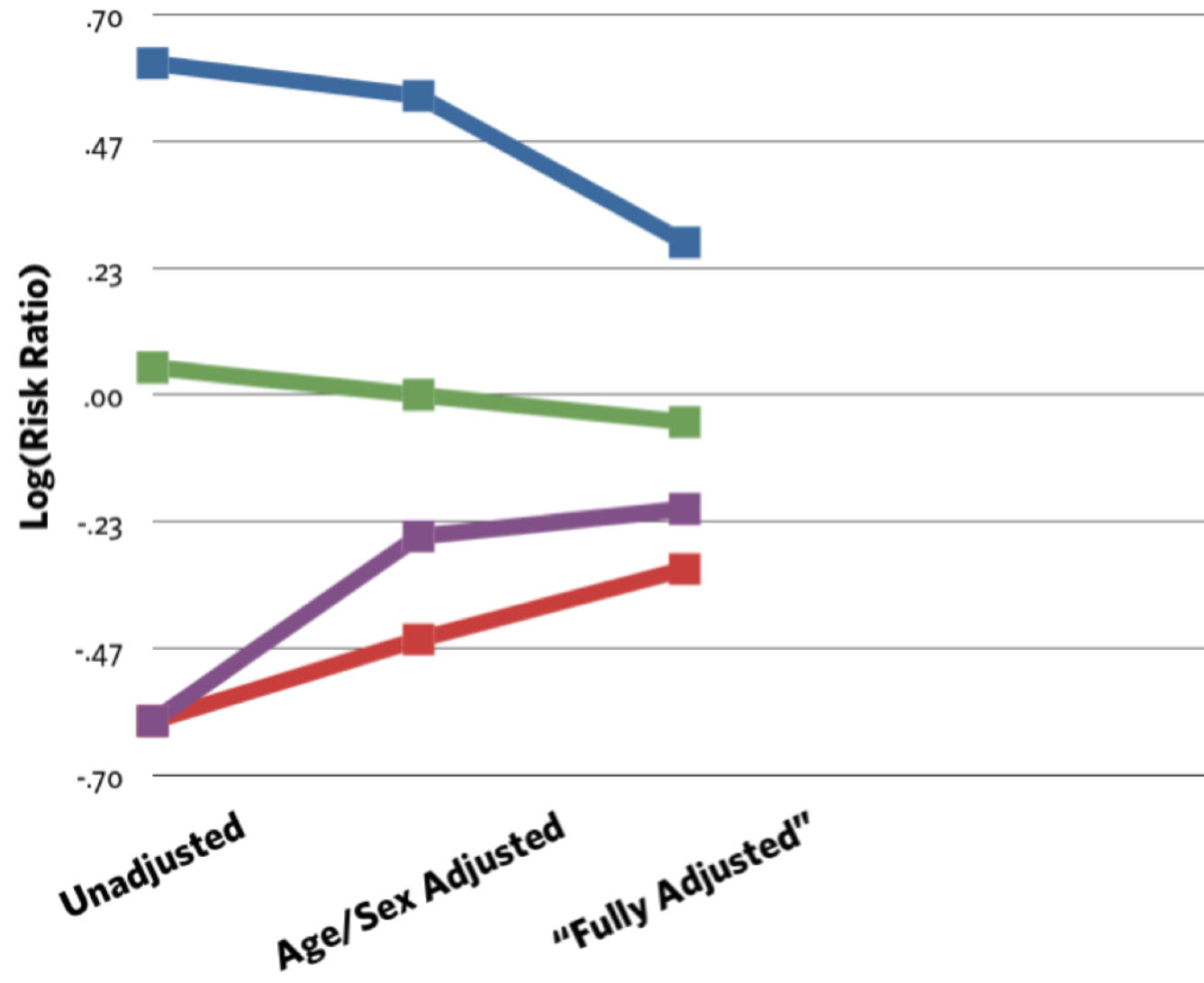
* OR = odds ratio; CI = confidence interval

Empirical results

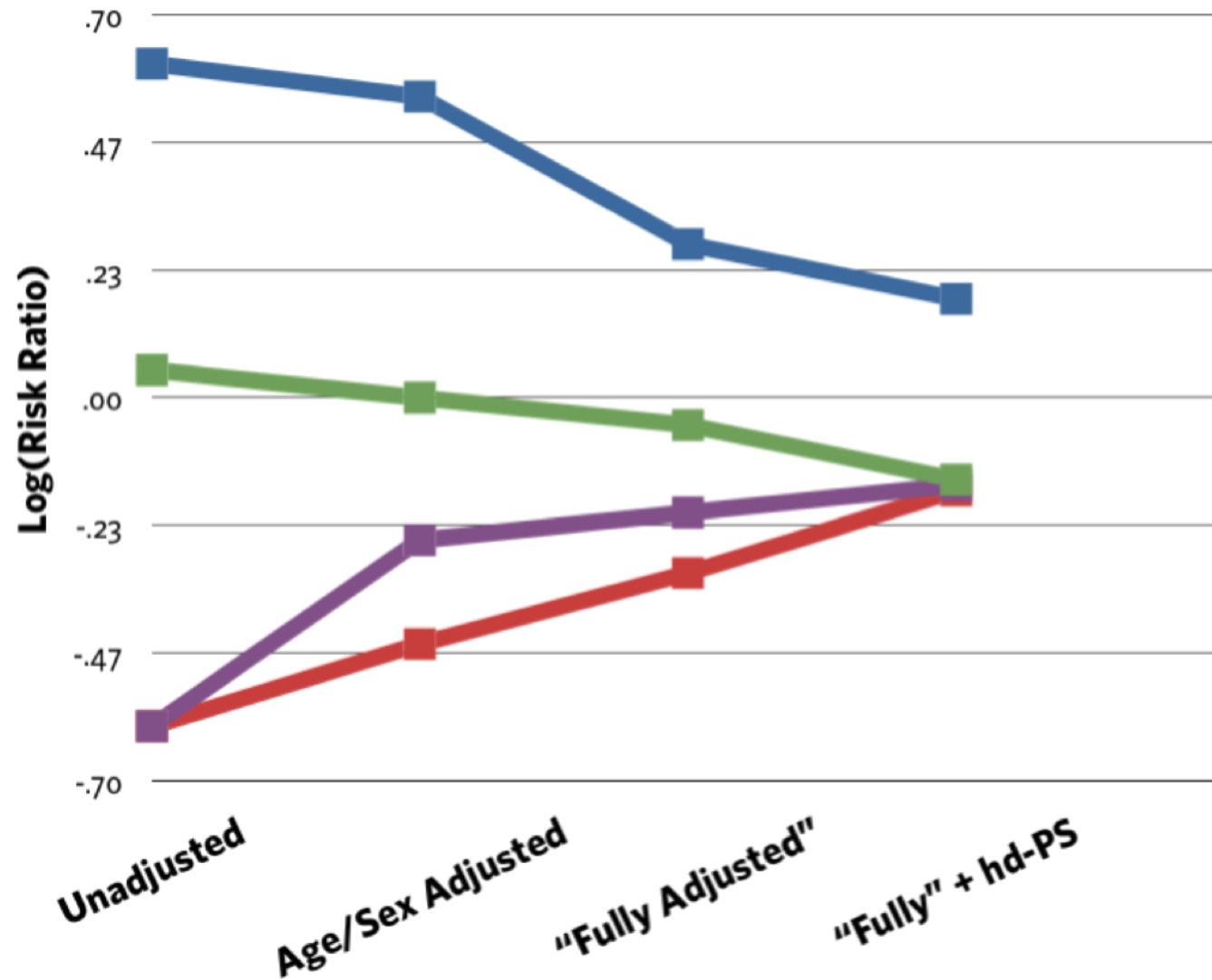
Table 3: Variations in covariate adjustment and relative risk estimates for the association of selective cox-2 inhibitors and GI complications within 180 days of first medication use.

Model #	Covariates included in propensity score model	Number of covariates adjusted	Variables tested per data source	Data source granularity	Covariate prioritization algorithm	c-statistic of PS model	Outcome model Relative risk	95% CI
N = 49,653								
1	Unadjusted					-	1.09	0.91-1.30
2	Age, sex, race, year**	d=4				0.61	1.01	0.84-1.21
3	+ predefined covars (Tab1)	d=4; l=14				0.66	0.94	0.78-1.12
4	+ empirical covariates	d=4; l=14; k=200	n=200	3-digit ICD	Bias _{mult}	0.69	0.86	0.72-1.04
5*	+ empirical covariates	d=4; l=14; k=500	n=200	3-digit ICD	Bias _{mult}	0.71	0.88	0.73-1.06
Bootstrapped 95% CIs:								0.73-1.06
5b	Only demographics + empirical covariates	d=4; k=500	n=200	3-digit ICD	Bias _{mult}	0.71	0.87	0.72-1.05

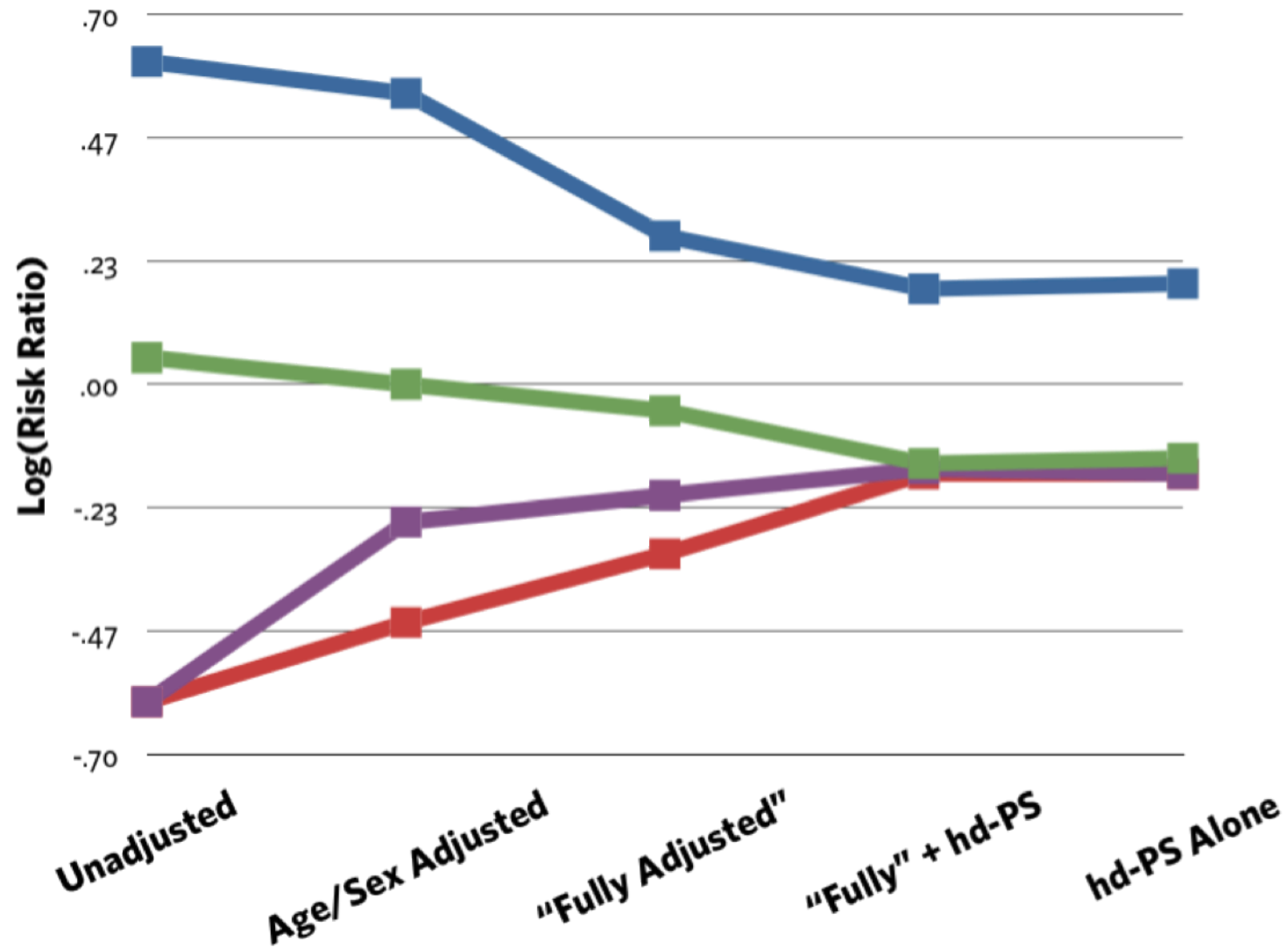
Recall the four studies



Recall the four studies



Recall the four studies



Conclusions

- ▶ Addressing confounding is paramount in CER
- ▶ Propensity scores are useful tools
- ▶ Leveraging the high-dimensional nature of administrative databases via propensity score techniques can improve confounding control
- ▶ hdPS code and documentation is available at drugepi.org/downloads

Thank you

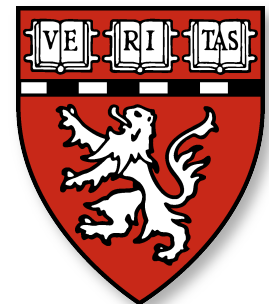
- ▶ Special thanks to Sebastian Schneeweiss, Jeremy Rassen, Robert Glynn, Jerry Avorn, Alan Brookhart, and Helen Mogun

Joshua Gagne (jgagne1@partners.org)

Division of Pharmacoepidemiology

Department of Medicine

Brigham and Women's Hospital



Additional slides

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

1	2	3	4	5	6	7	8
MC Part A* Hosp Dx (ICD-9)	MC Part A Hosp proc (ICD-9)	MC Part A Amb Dx (ICD-9)	MC Part A Amb proc (ICD-9)	MC Part A Nurse Home (ICD-9)	Part B Dx (ICD-9)	Part B Proc (CPT-4)	Drugs generic entities

* MC = Medicare, Hosp = hospital, Amb = ambulatory, Dx = diagnosis, proc = procedure, Rx = prescription dispensings

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

hd-PS algorithm (3/7)

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

For each identified code create three covariates for each patient:

- ▶ `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

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:

$$\text{Bias}_{\text{mult}} = \frac{P_{\text{CI}}(RR_{\text{CI}} - 1) + 1}{P_{\text{CO}}(RR_{\text{CO}} - 1) + 1}$$

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 exposure 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 in outcome model(s)

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

Some first observations (1)

- ▶ In our examples...
- ▶ Adding the hd-PS to the predefined covariates moved the point estimate in the expected direction
- ▶ Results of the hd-PS alone were identical to the second decimal compared with the hd-PS combined plus predefined covariates
- ▶ Including 500 rather than 200 covariates into the hd-PS appeared to move the effect estimate slightly in the expected direction

Some first observations (2)

- ▶ More granular diagnostic codes (4 digit ICD-9 vs. 3) appeared to move the effect estimate slightly but not substantially in the expected direction
- ▶ Recurrence assessment seems important (ie, code appeared once, twice, ... times)

hd-PS in some (even) tougher
situations

Algorithm meets world

- ▶ hd-PS is data-hungry
- ▶ What happens when hd-PS has...
 - ▶ Small sample sizes?
 - ▶ Few exposures and/or outcomes?
- ▶ What happens to hd-PS when...
 - ▶ We decrease the number of selected covariates?

Three example studies

- ▶ STUDY 1

Exposure: coxibs vs. ns-NSAIDS

Outcome: GI bleed

Population: 49,653 low income, older adults in Pennsylvania

- ▶ STUDY 2

Exposure: statins vs. glaucoma drugs

Outcome: all-cause mortality

Population: 36,122 low income, older adults in Pennsylvania

- ▶ STUDY 3

Exposure: statins vs. glaucoma drugs

Outcome: MI or cardiovascular death

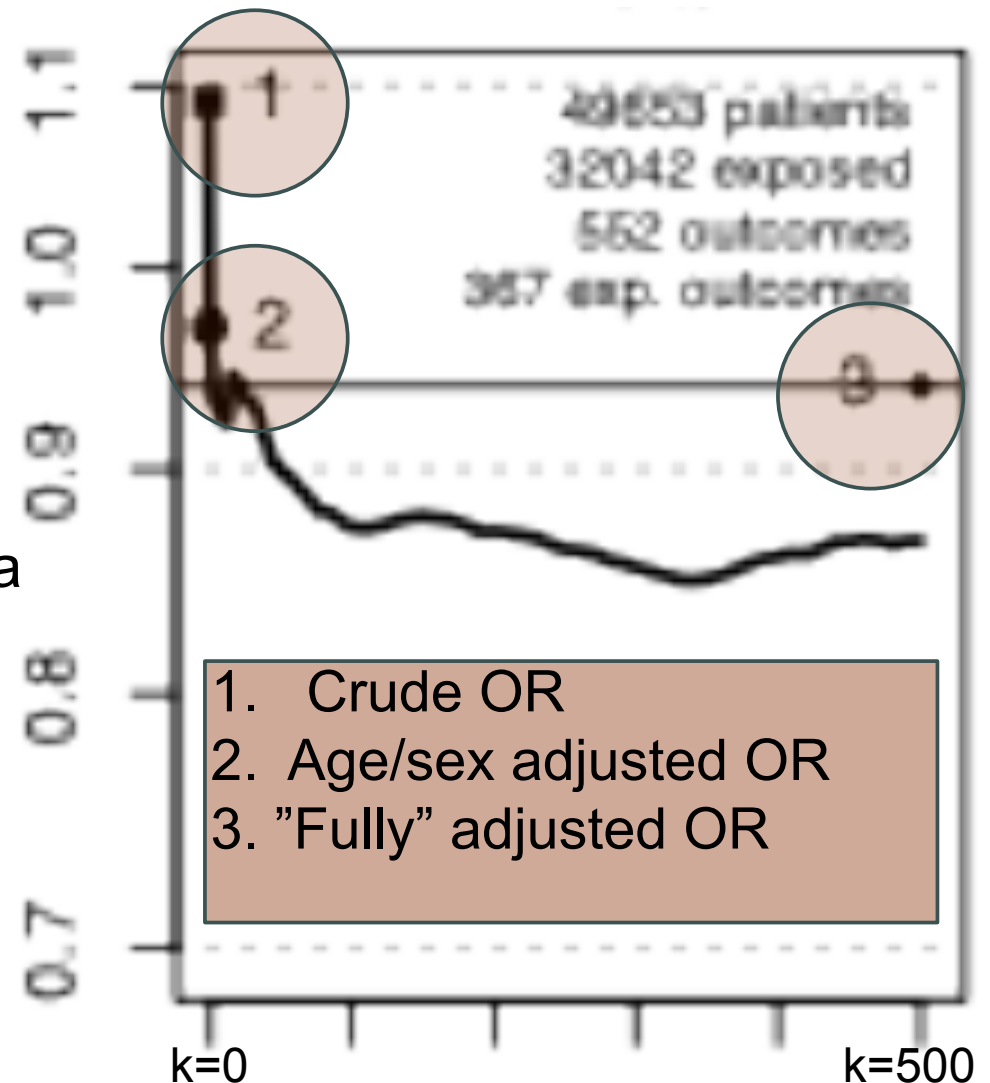
Population: 36,122 low income, older adults in Pennsylvania

Test the abilities of hd-PS in boundary conditions

- ▶ For each of three example studies:
 - ▶ Sample the original study population at rates of 5%, 10%, 15%, 20%, and 50%.
 - ▶ Run the hd-PS algorithm with $k=0$ to $k=500$ empirically-identified covariates.
 - ▶ Record observed odds ratio (OR) at each level of k and several performance metrics.
 - ▶ Repeat 50 times.

Example notation and graph

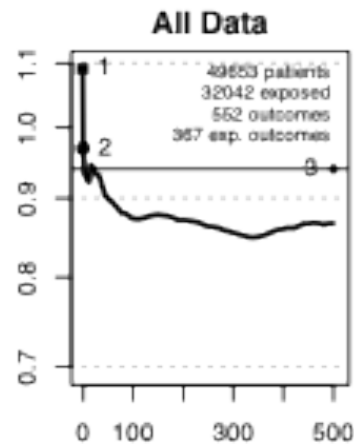
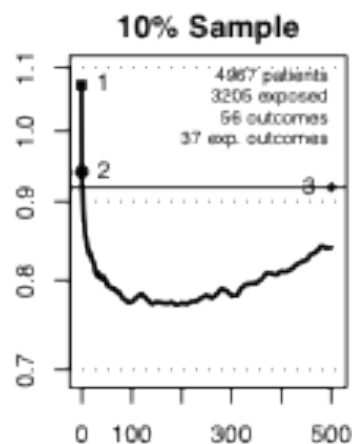
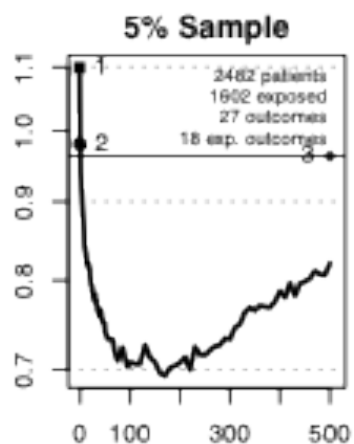
- ▶ Exposure:
coxibs vs. ns-NSAIDS
- ▶ Outcome:
GI bleed
- ▶ Population:
49,653 low income,
older adults in Pennsylvania



Outcome measures

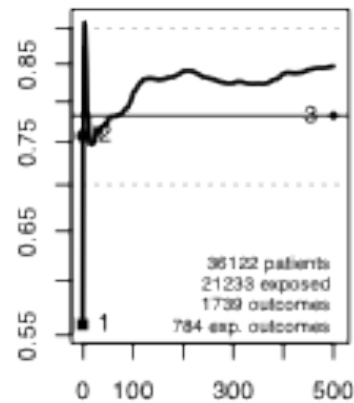
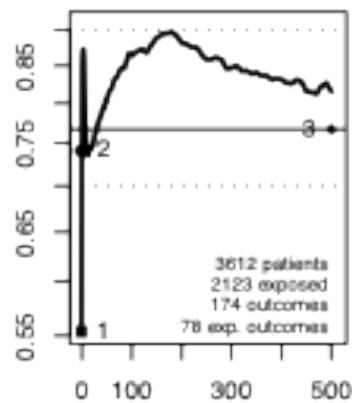
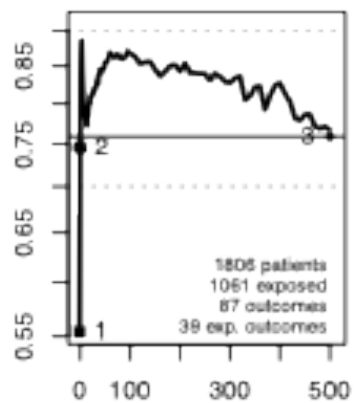
- ▶ Mean odds ratio at each level of k
- ▶ Odds ratio difference: mean percentage difference between the sample's observed OR and the OR in the full cohort (lower is better).
“Did we get the right result?”
- ▶ Variable coverage: mean percentage of variables identified in the sample that were also identified in the full sample (higher is better).
“Did we pick the right variables?”

Coxib /
GI Bleed

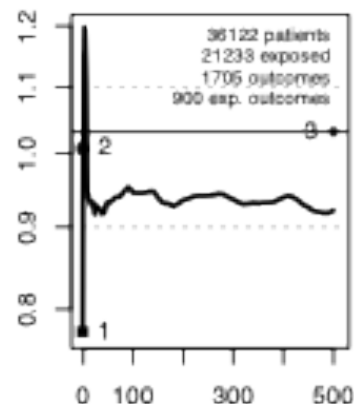
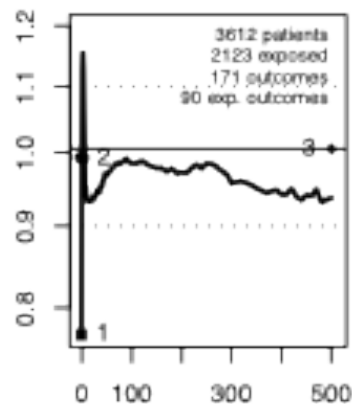
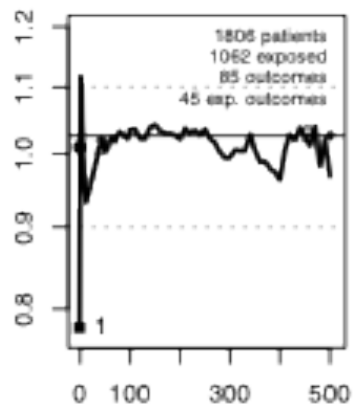


1=crude
2=age/sex adjusted
3=fully adjusted
(solid line)

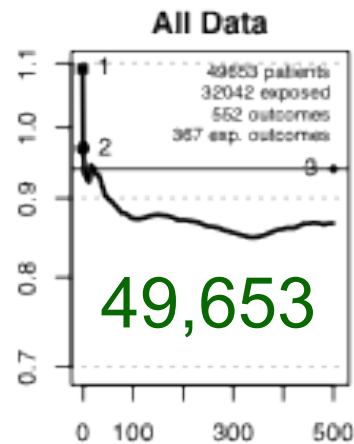
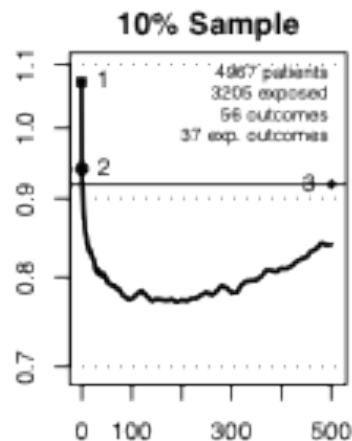
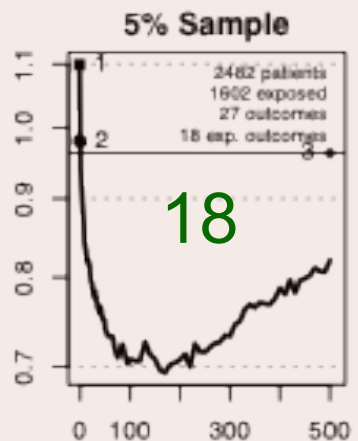
Statin /
Death



Statin /
MI or
CV Dth

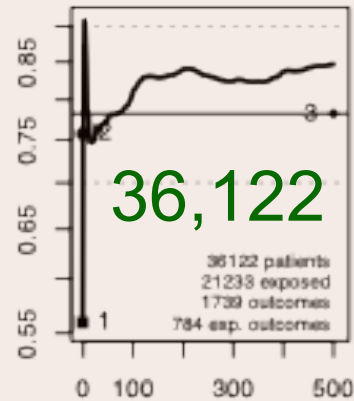
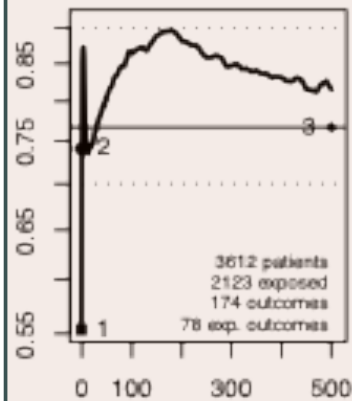
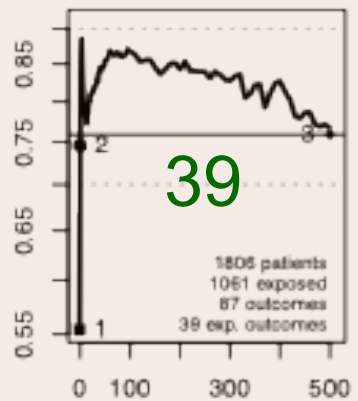


Few
exposed
outcomes.

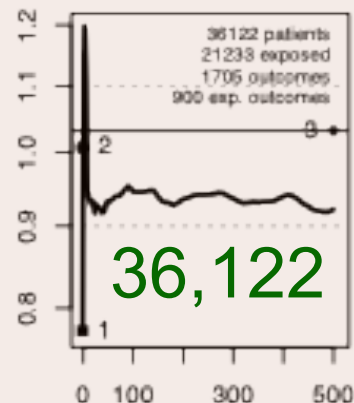
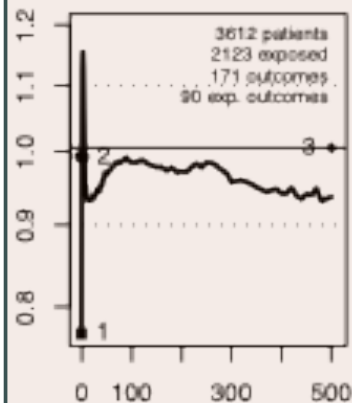
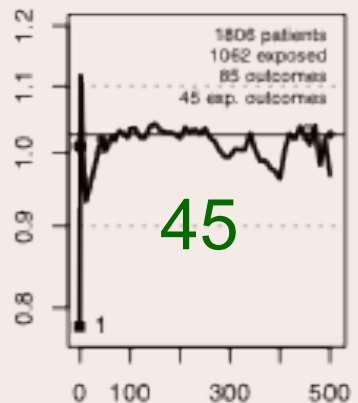


1=crude
2=age/sex adjusted
3=fully adjusted
(solid line)

More
exposed
outcomes.



Approaches
“true” OR
in sample
at low value
of k, even
though these
are smaller
studies overall.



Odds Ratio Difference

Average percentage difference between the sample's observed OR and the OR in the full cohort (lower is better).

	COXIB	STATIN AND DEATH	STATIN AND MI OR CV DEATH
"5% SAMPLE"	43%	69%	58%
"10% SAMPLE"	25%	22%	15%
"20% SAMPLE"	51%	7%	44%
"50% SAMPLE"	8%	4%	7%

1. General downward trend.
2. There exists a number at which good performance is reached.

Variable Coverage

Average percentage of variables identified in the small sample that were also identified in the full sample (higher is better).

	COXIB	STATIN AND DEATH	STATIN AND MI OR CV DEATH
“5% SAMPLE”	55%	59%	58%
“10% SAMPLE”	59%	65%	63%
“20% SAMPLE”	63%	73%	70%
“50% SAMPLE”	72%	74%	82%

Very good OR performance correlates with just 72% to 82% variable performance: last variables added don't make a big difference.

Some more observations

- ▶ In the full cohort, at approximately $k=250$ empirically-identified covariates, the maximal level of confounding adjustment was reached. (More didn't hurt.*)
- ▶ At low values of k , the observed OR was very sensitive to addition of new variables. (Strong confounders first.)
- ▶ hd-PS did not perform optimally with small numbers of exposed outcomes (~150-200 were needed).
- ▶ In all cases, the hd-PS adjusted as well as or better than conventionally-adjusted studies

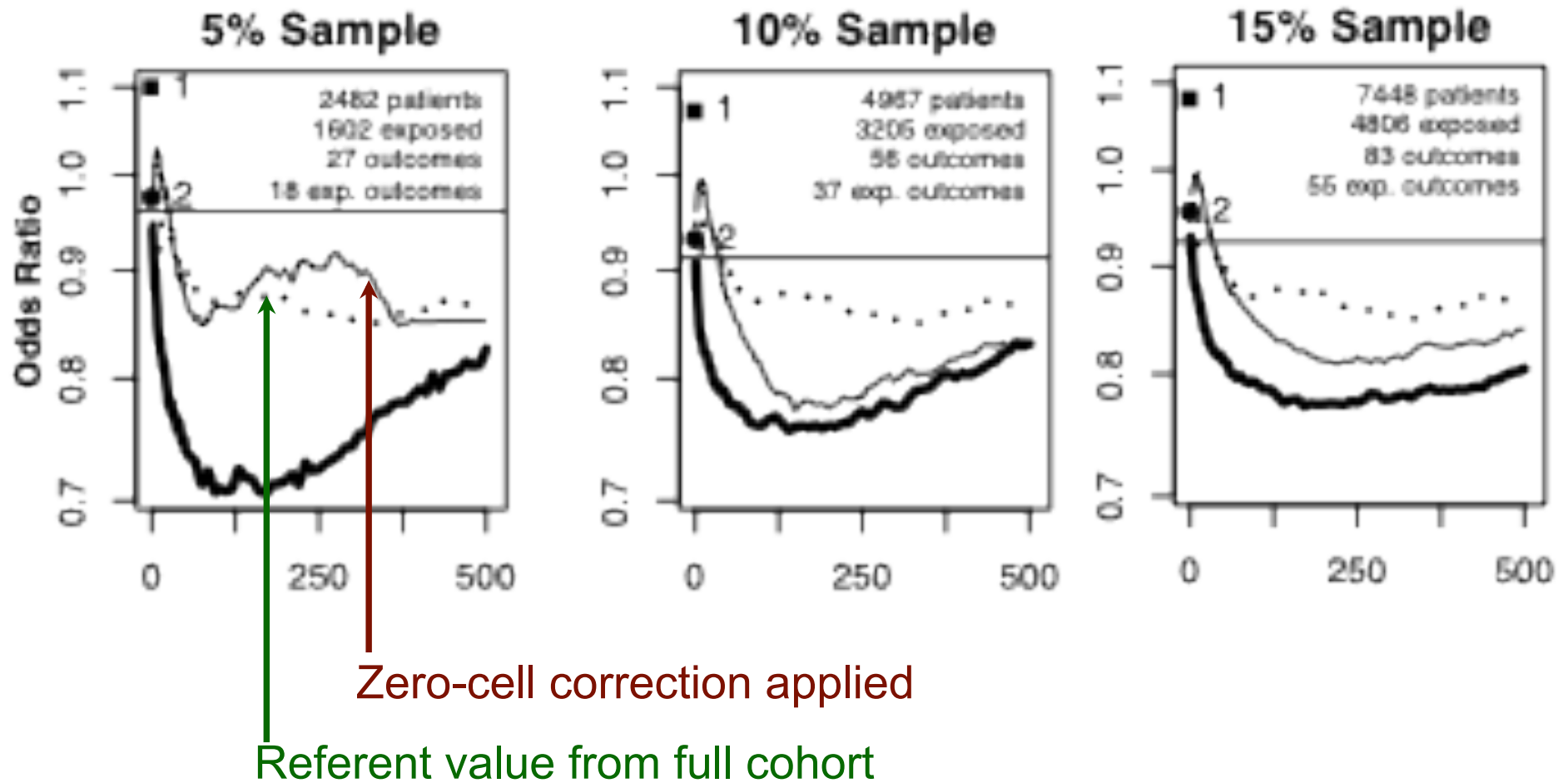
And some fixes

- ▶ One challenge with few outcomes is that the algorithm rules out many potentially-useful confounders because of zeros in the confounder -> outcome 2x2 table:

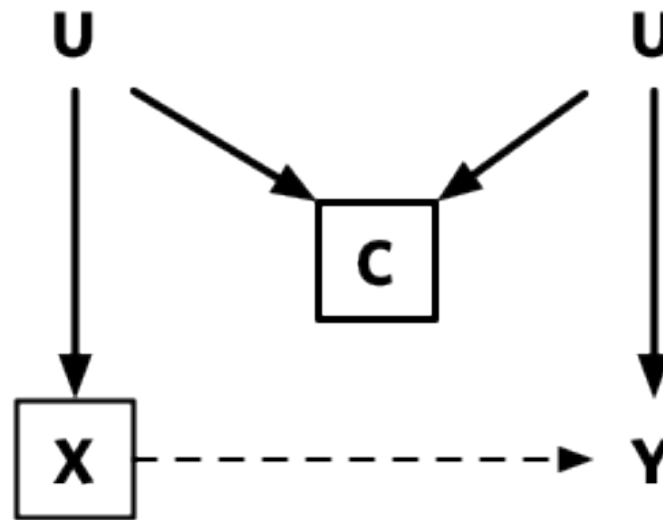
	CONF.	NO CONF.
OUTCOME	2	0
NO OUTCOME	9,423	8,805

- ▶ Solution 1: screen only on confounder -> exposure relationship
- ▶ Solution 2: add a small correction factor (say, 0.1) to each cell.

Example of fixes: Coxib/GI Bleed

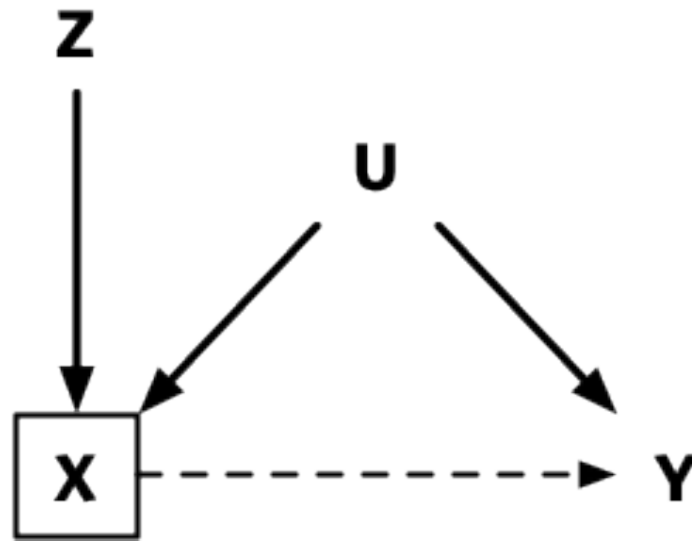


Collider-stratification bias Type 1: M-bias



- M-bias is possible but usually considered weak.

Collider-stratification bias Type 2: Z-bias



- Conditioning on treatment will open a backdoor path such that an instrument-like variable (which was not originally a confounder) can cause bias.

Greenland, Epidemiology
2003.

Brookhart, AJE 2006.

Diagnostics for Z-bias (coxib example)

Dimension	code	frequency_ type	rr_ce	rr_cd	rr_ce_ rankStars	
x3_ambdx	274	once	0.57413	1.33750	1.0	***
dx3_mddx	274	once	0.59632	1.55786	2.0	*
dx3_hospdx	618	once	0.68079	1.29027	3.0	***
dx3_ambdx	618	once	0.68825	1.20109	4.0	****
dx3_ambdx	162	sporadic	0.75304	1.23305	6.0	****
generic_drugs	warfarin sodium	frequent	1.29333	1.33805	12.0	***
dx3_nhdx	715	frequent	1.28900	1.34443	14.0	***
prcdr_mdproc	85610	sporadic	1.28382	1.56258	16.0	*
prcdr_ambproc	863	once	0.77982	1.27691	17.0	****
generic_drugs	warfarin sodium	sporadic	1.28004	1.34342	20.0	Arthritis
dx3_mddx	714	frequent	1.27896	1.05056	21.0	*****
dx3_ambdx	714	frequent	1.27299	1.46914	23.0	**
generic_drugs	warfarin sodium	once	1.26605	1.39066	25.0	**
dx3_mddx	714	sporadic	1.26183	1.14412	27.0	*****
dx3_ambdx	162	once	0.79712	1.14912	31.0	****
dx3_mddx	725	once	1.24455	1.15490	35.0	****
generic_drugs	olanzapine	frequent	0.80962	1.46871	38.0	**
generic_drugs	folic acid	frequent	1.23513	1.52650	39.0	*
prcdr_mdproc	85610	once	1.22844	1.38142	44.0	***
dx3_hospdx	332	frequent	1.22719	1.24977	45.0	****
generic_drugs	tramadol hcl	sporadic	1.22716	1.46824	46.0	**
dx3_nhdx	427	frequent	1.22433	1.58125	50.0	*
dx3_mddx	714	once	1.22422	1.22965	51.0	****
generic_drugs	calcitonin, salmon, synthetic	sporadic	1.22363	1.22610	52.0	****
dx3_nhdx	V43	once	1.22324	1.21224	53.5	****
generic_drugs	calcitonin, salmon, synthetic	frequent	1.22253	1.50926	55.0	*
dx3_mddx	715	once	1.22232	1.06860	56.0	*****

Key hd-PS limitations

- ▶ All corner cases are hard to test, as it is difficult to simulate high dimensional data.
- ▶ Including many variables increases the chance of M- or Z-bias; diagnostics are key.
- ▶ Somewhat less transparent than standard adjustment techniques.

Available now from drugepi.org

- ▶ hd-PS Version 1
 - ▶ The original SAS macro
- ▶ hd-PS Version 2 (beta)
 - ▶ (Much) faster running times
 - ▶ Zero-cell correction and other improvements
 - ▶ Java-based implementation with full SAS macro
 - ▶ R template code
 - ▶ Ability to use in other systems (surveillance, etc.)

Coming soon

- ▶ hdpharmacoepi.org web site
 - ▶ SAS and R code generation
 - ▶ Study results repository
 - ▶ Automated table generation
 - ▶ Automated code lookup
 - ▶ Additional diagnostics

hd-PS conclusions

- ▶ We believe that in many instances, hd-PS may perform better than conventional confounding adjustment.
- ▶ It performs as well as or better than investigator-defined covariates
- ▶ It reduces potential for error in covariate definition
- ▶ It is fast to implement
- ▶ It can be easily applied to subgroup analyses.