OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Observational study design

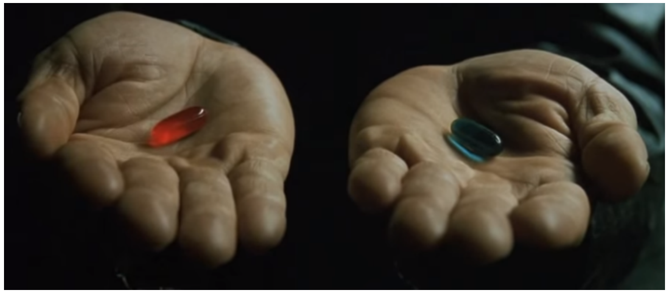
Patrick Ryan, PhD

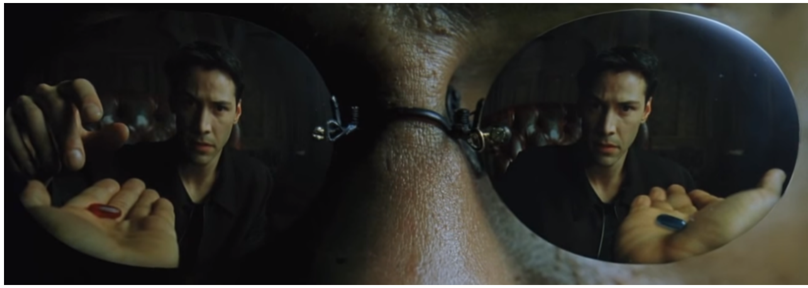
Columbia University

Janssen Research and Development

A little exercise:
choose your own adventure!

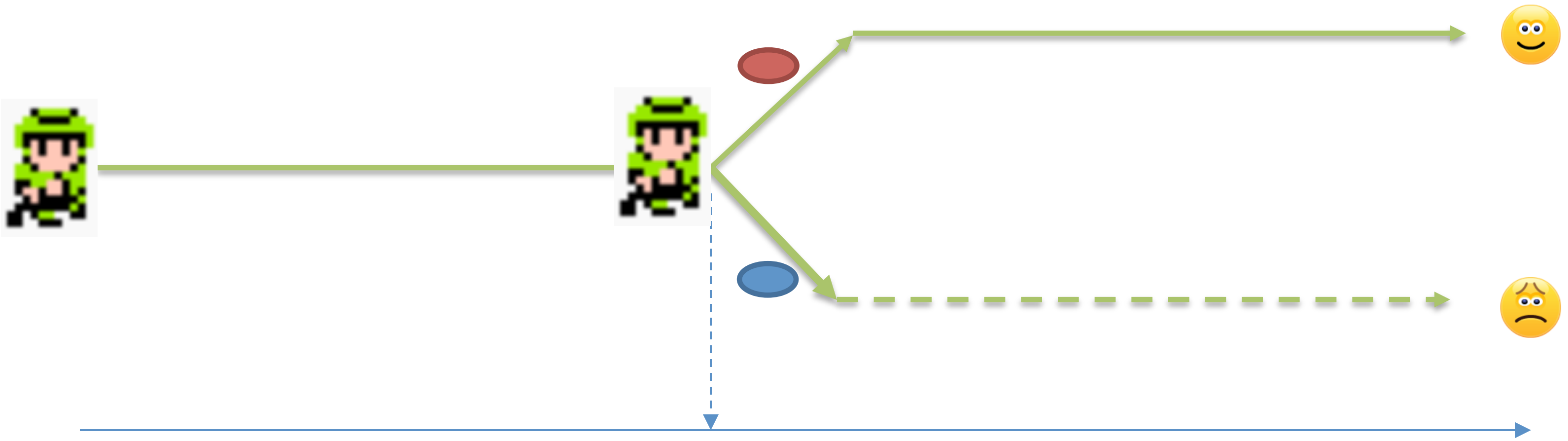
A pop culture mash-up to explain counterfactual reasoning…





Counterfactual reasoning for one person

Decision



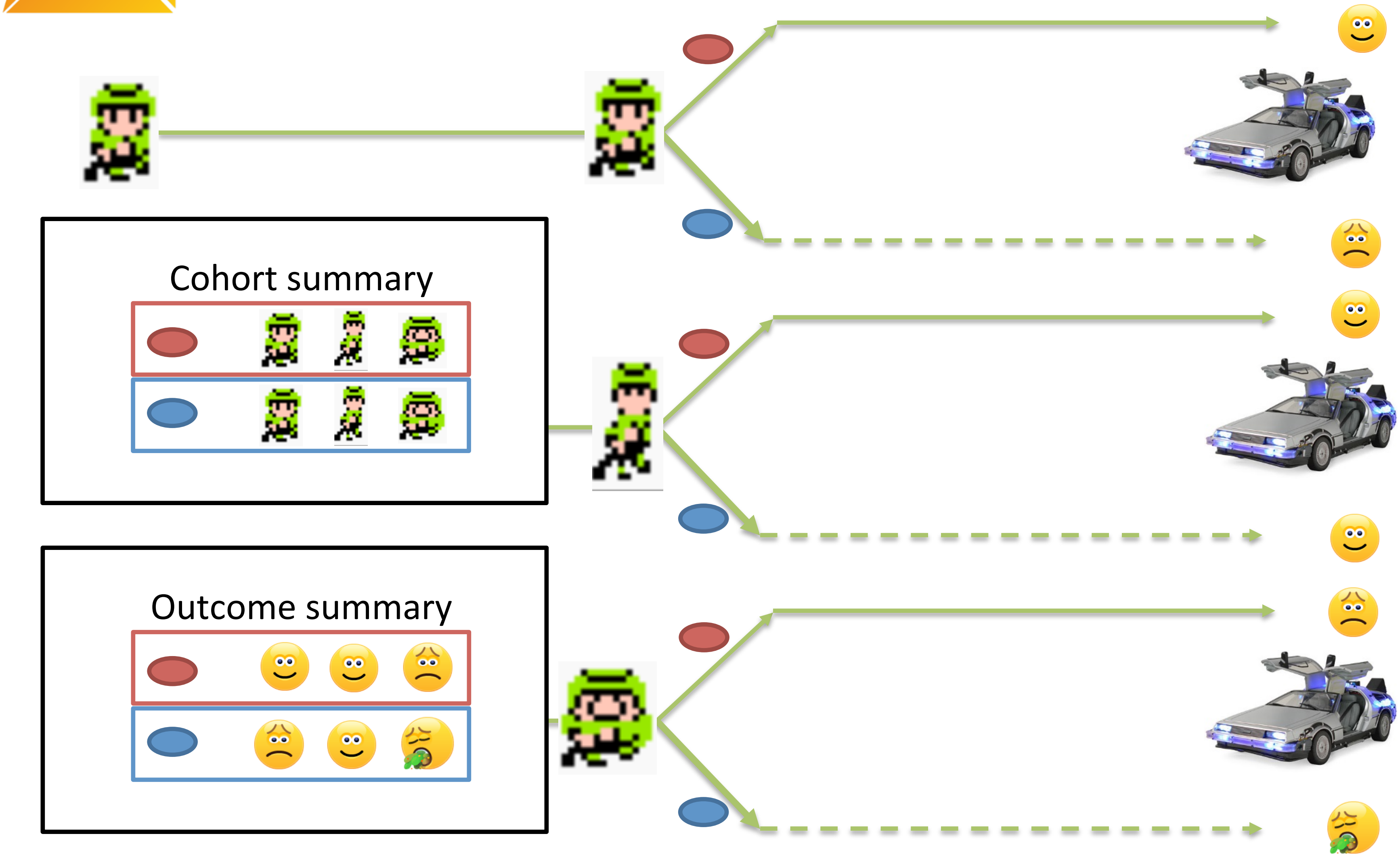
Baseline:
Period to satisfy
inclusion criteria

0

Follow-up:
Period to observe
outcomes

Person Time

Counterfactual reasoning for a population



Cohort summary

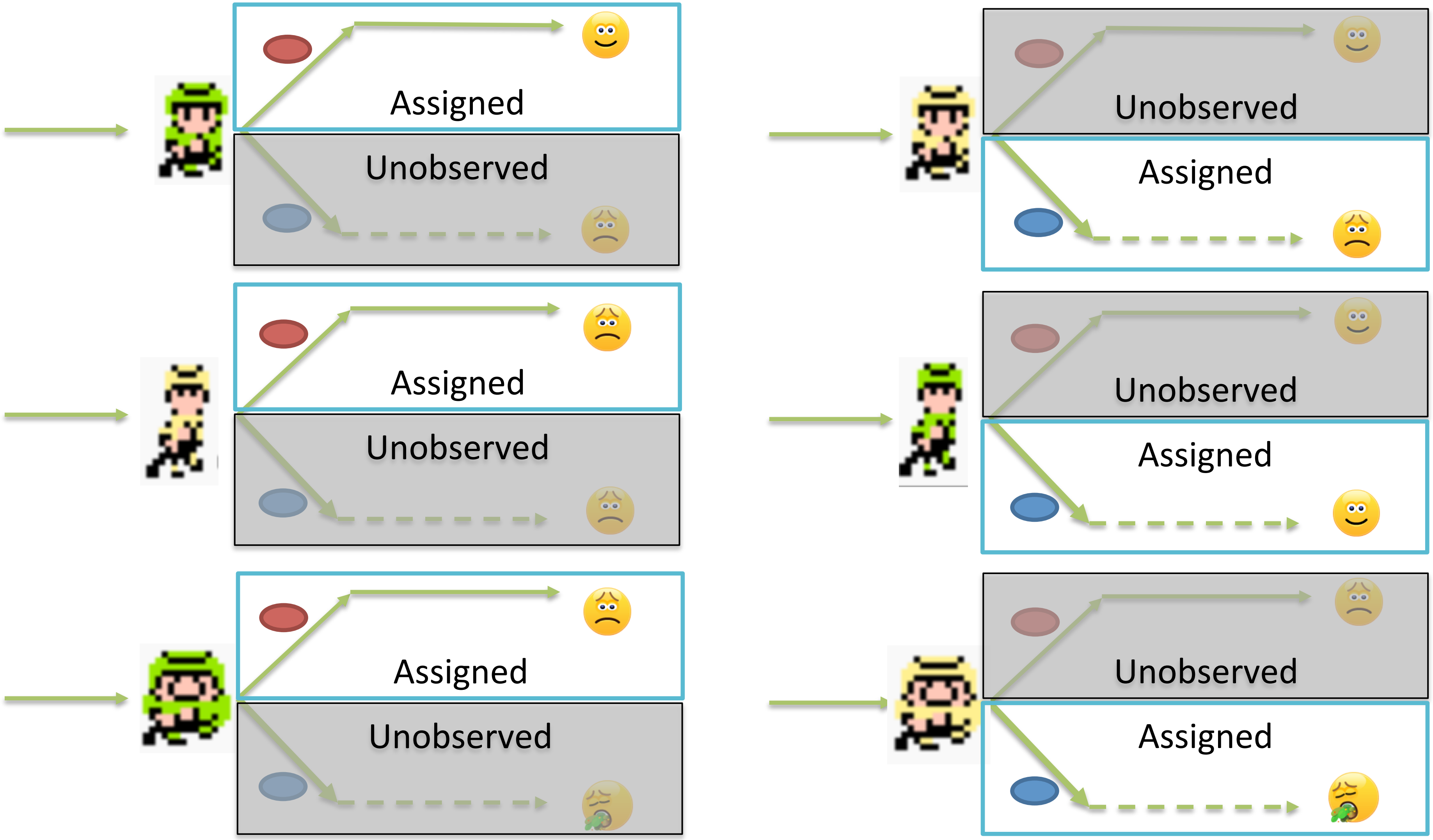
Outcome summary

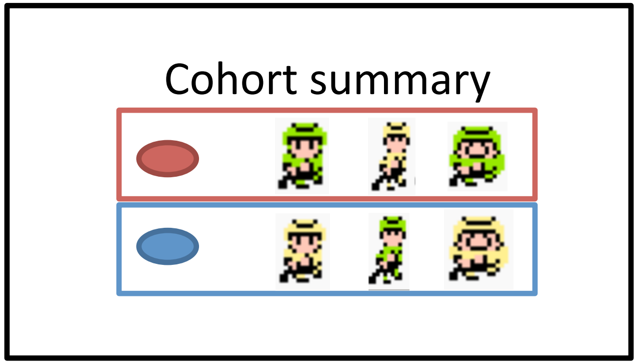
Alas, we don't have a Delorean…

- What is our *next* best approximation?

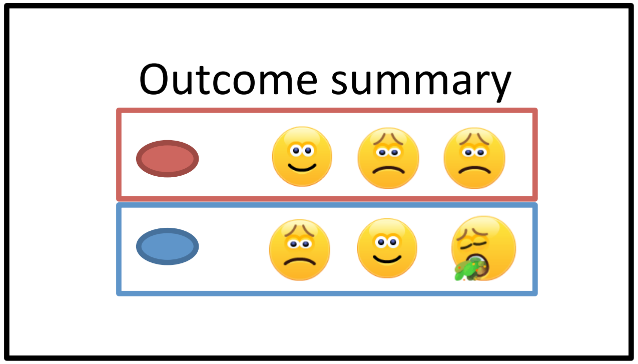
- Instead of studying the same population under both decision options, let's define a larger population and randomly assign one treatment to each person, then compare outcomes between the two cohorts…

Randomized treatment assignment to approximate counterfactual outcomes





Cohort summary



Outcome summary

- Randomization allows for assumption that persons assigned to target cohort are exchangeable at baseline with persons assigned to comparator cohort

Alas, we can't randomize…

- What is our *next, next* best approximation?

- Define a larger population, observe the treatment choices that were made, then compare outcomes:
  - Between persons who made different choices (comparative cohort design)

  OR

  - Within persons during time periods with different exposure status (self-controlled designs)

How does Epidemiology define a comparative cohort study?

…it depends on what Epidemiology textbook you read…

“In a retrospective cohort study…the investigator identified the cohort of
individu
their su
recent

“Cohort studies are studies that identify subsets of a defined population and follow them over time, looking for differences in their outcome. Cohort studies generally compare exposed patients to unexposed patients, although they can also be used to compare one exposure to another.”

--Strom, Pharmacoepidemiology, 2005

“In
has experienced the outcome of interest, but all of whom could experience it…
On en
charac
people

“In the coho
identified. I
incidence of
ascertained
--S

“In the paradigmatic cohort study, the investigator defines two or more groups of people that are free of disease and that differ according to the extent of their exposure to a potential cause of disease. These groups are referred to as the study cohorts. When two groups are studies, one is usually though of as the exposed or index cohort – those individuals who have experienced the putative causal event or condition – and the other is then thought of as the unexposed or reference cohort.”

--Rothman, Modern Epidemiology, 2008

OHDSI’s definition of ‘cohort’

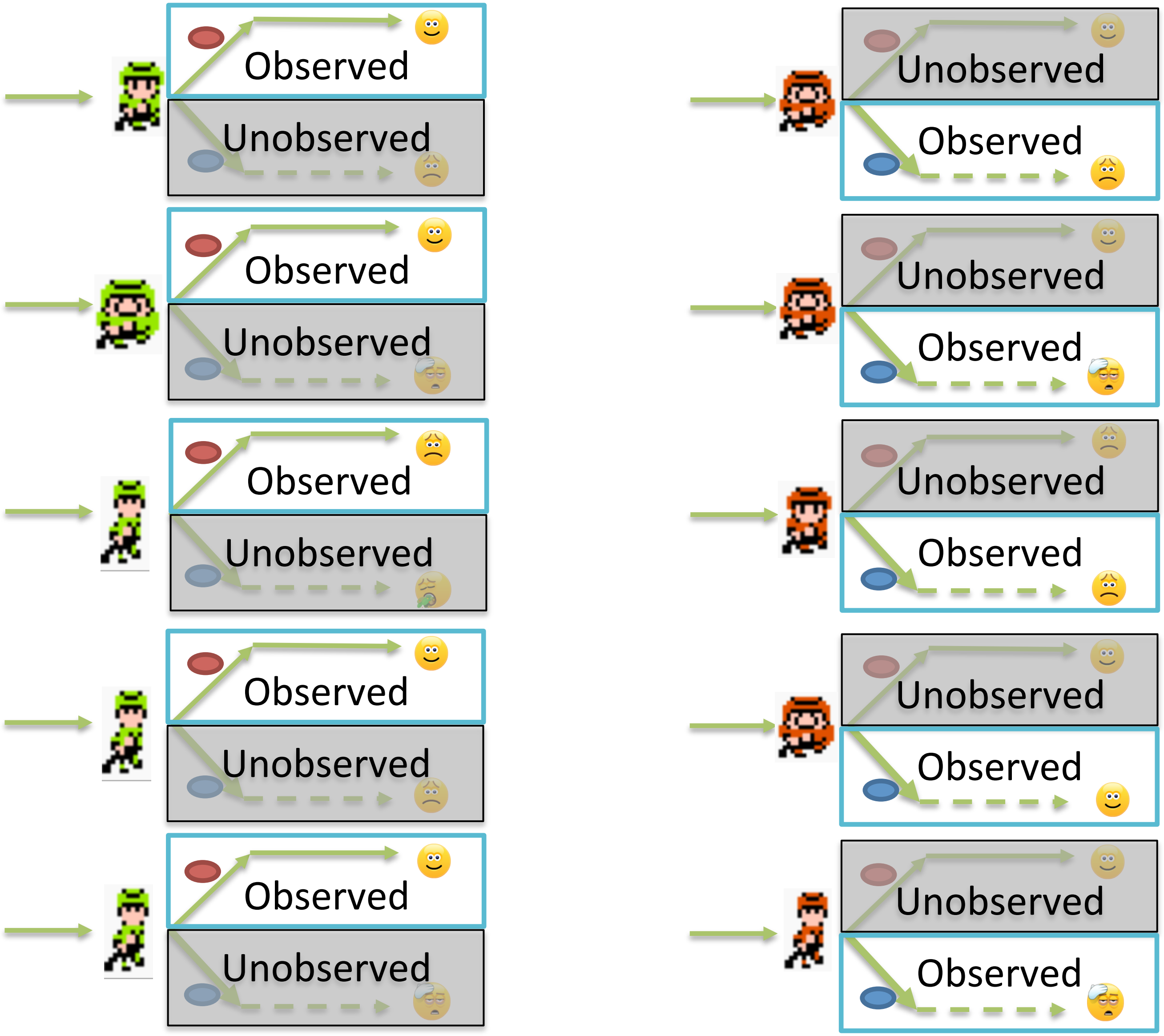
Cohort = a set of persons who satisfy one or more inclusion criteria for a duration of time

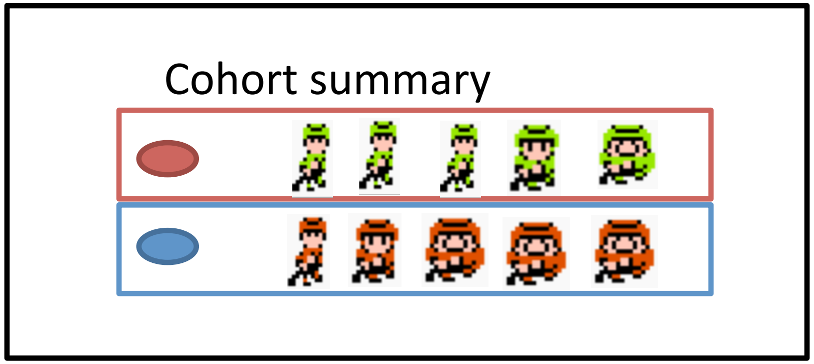
Objective consequences based on this cohort definition:

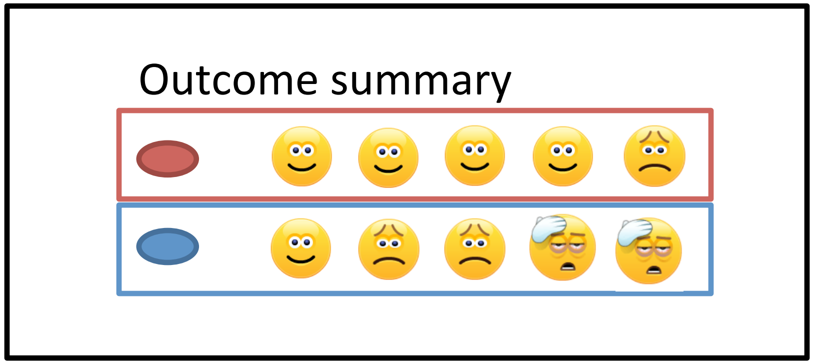
- One person may belong to multiple cohorts
- One person may belong to the same cohort at multiple different time periods
- One person may not belong to the same cohort multiple times during the same period of time
- One cohort may have zero or more members
- A codeset is NOT a cohort…

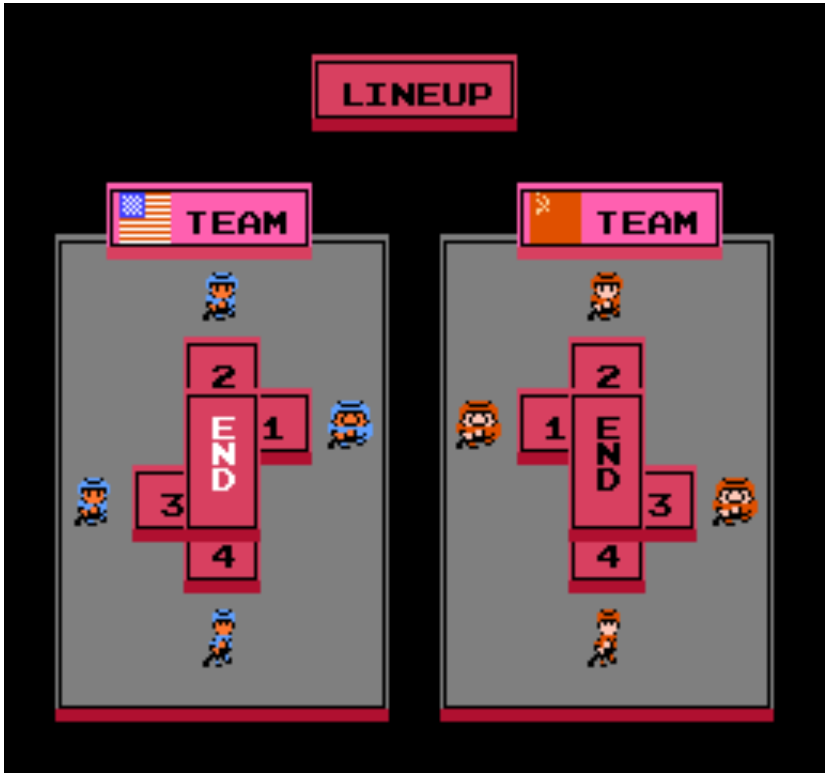
…logic for how to use the codeset in a criteria is required

An observational comparative cohort design to approximate counterfactual outcomes









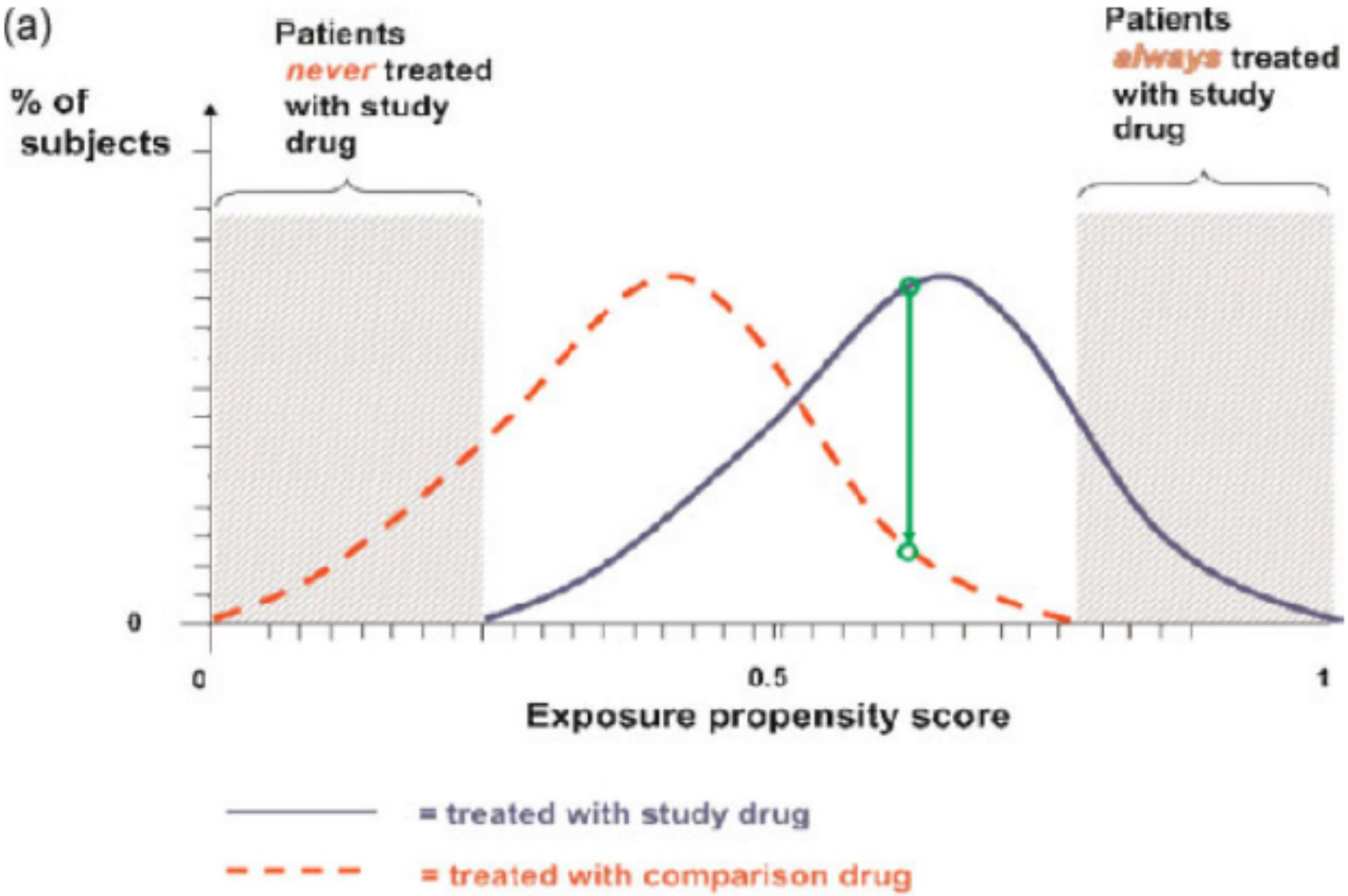
- Exchangeability assumption may be violated if there is reason for treatment choice...and there often is

Propensity score introduction

- $e(x) = Pr(Z=1|x)$
  - Z is treatment assignment
  - x is a set of all covariates at the time of treatment assignment
- Propensity score = probability of belonging to the target cohort vs. the comparator cohort, given the baseline covariates
- Propensity score can be used as a 'balancing score': if the two cohorts have similar propensity score distribution, then the distribution of covariates should be the similar (need to perform diagnostic to check)

Rubin Biometrika 1983

Intuition around propensity score balance



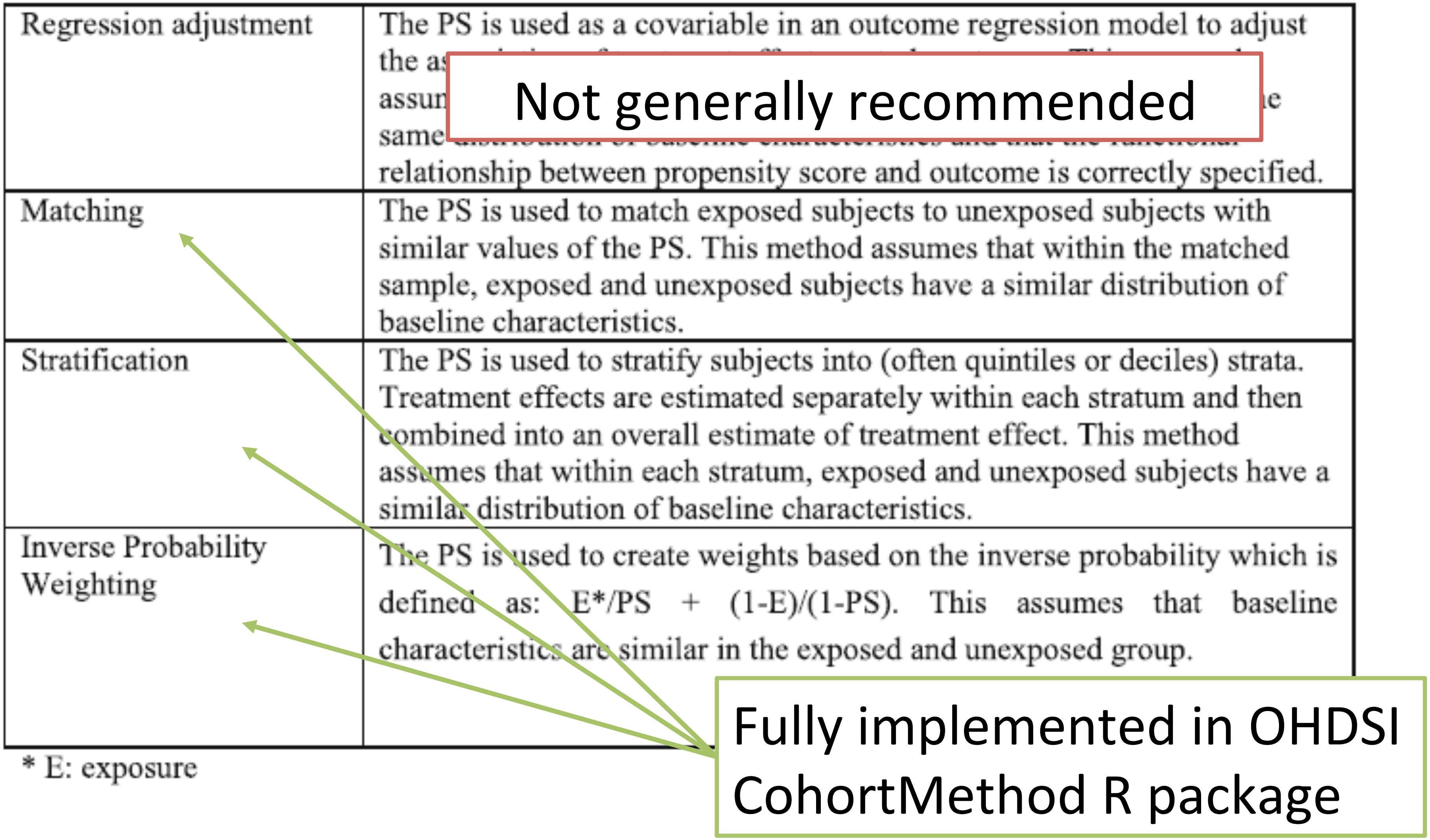
Schneeweiss. PDS 2011

“Five reasons to use propensity score in pharmacoepidemiology”

- Theoretical advantages
  - Confounding by indication is the primary threat to validity, PS focuses directly on indications for use and non-use of drug under study
- Value of propensity scores for matching or trimming the population
  - Eliminate ‘uncomparable’ controls without assumptions of linear relationship between PS and outcome
- Improved estimation with few outcomes
  - PS allows matching on one scalar value rather than needing degrees of freedom for all covariates
- Propensity score by treatment interactions
  - PS enables exploration of patient-level heterogeneity in response
- Propensity score calibration to correct for measurement error

Glynn et al, BCPT 2006

Methods for confounding adjustment using a propensity score



Regression adjustment	The PS is used as a covariable in an outcome regression model to adjust the as assu ne same relationship between propensity score and outcome is correctly specified.
Matching	The PS is used to match exposed subjects to unexposed subjects with similar values of the PS. This method assumes that within the matched sample, exposed and unexposed subjects have a similar distribution of baseline characteristics.
Stratification	The PS is used to stratify subjects into (often quintiles or deciles) strata. Treatment effects are estimated separately within each stratum and then combined into an overall estimate of treatment effect. This method assumes that within each stratum, exposed and unexposed subjects have a similar distribution of baseline characteristics.
Inverse Probability Weighting	The PS is used to create weights based on the inverse probability which is defined as: E*/PS + (1-E)/(1-PS). This assumes that baseline characteristics are similar in the exposed and unexposed group.

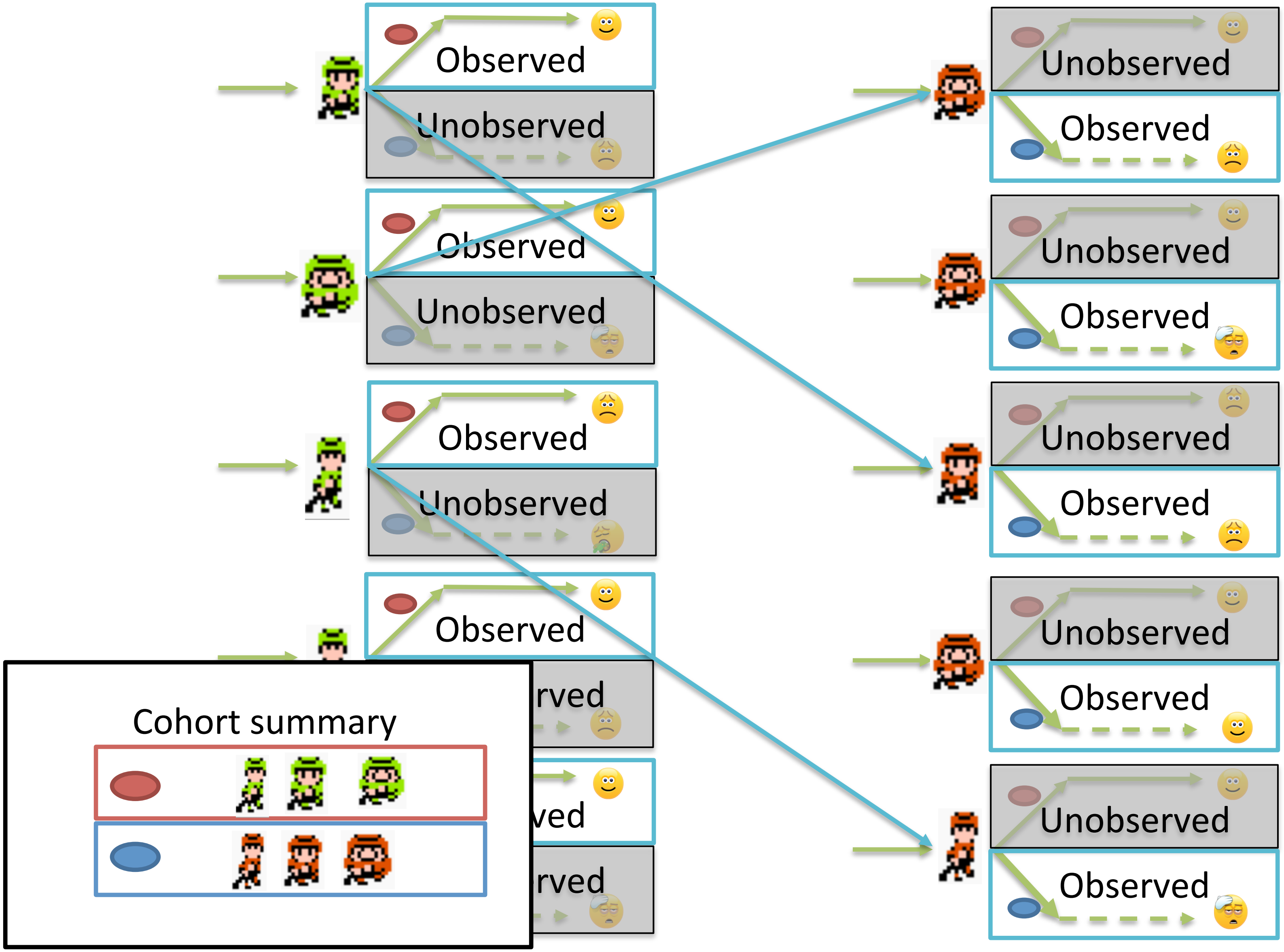
* E: exposure

Not generally recommended

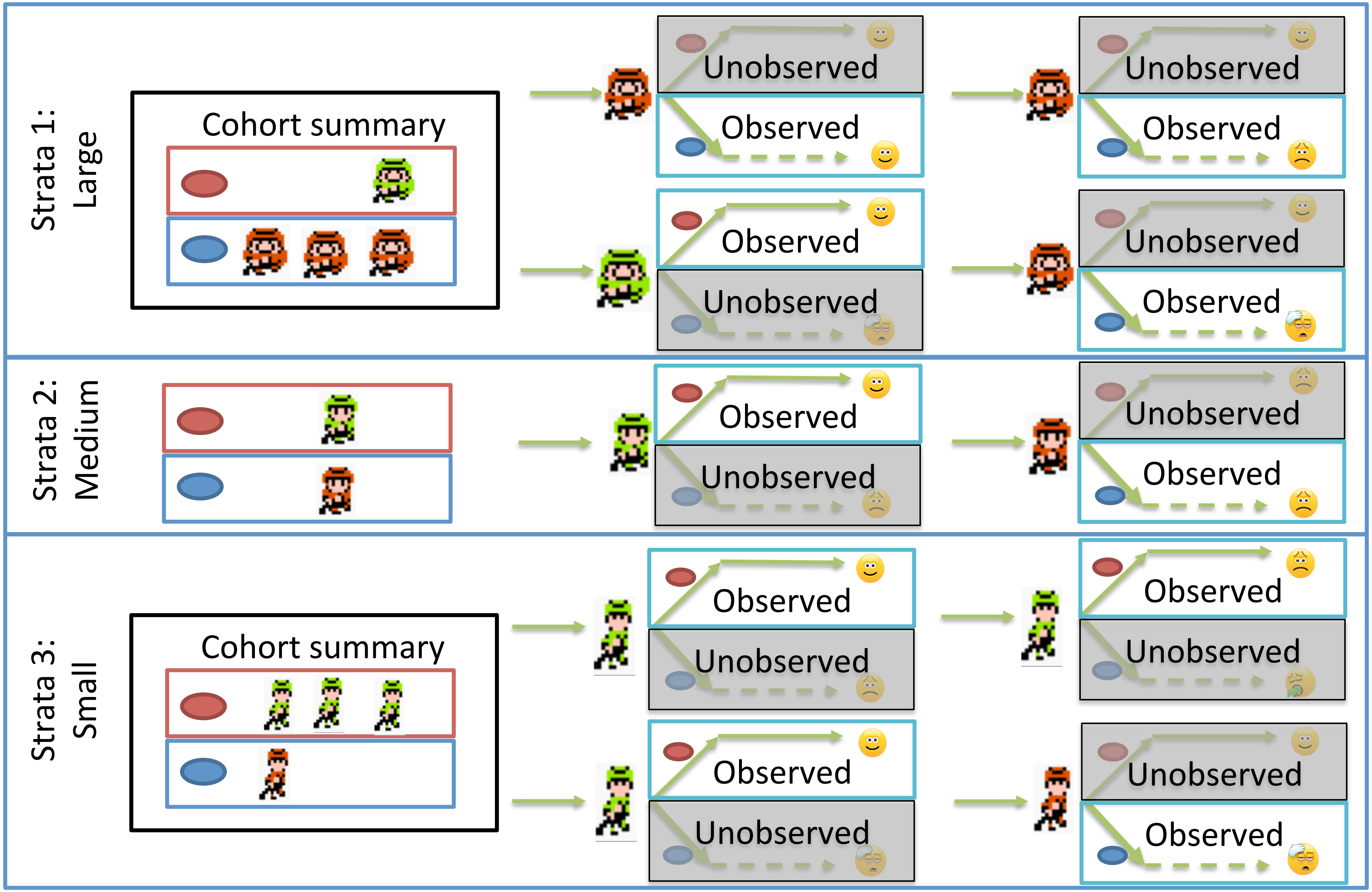
Fully implemented in OHDSI CohortMethod R package

Garbe et al, Eur J Clin Pharmacol 2013, http://www.ncbi.nlm.nih.gov/pubmed/22763756

Matching as a strategy to adjust for baseline covariate imbalance



Stratification as a strategy to adjust for baseline covariate imbalance



Strata 1: Large

Cohort summary

Unobserved

Observed

Unobserved

Observed

Observed

Unobserved

Unobserved

Observed

Strata 2: Medium

Observed

Unobserved

Unobserved

Observed

Strata 3: Small

Cohort summary

Observed

Unobserved

Observed

Unobserved

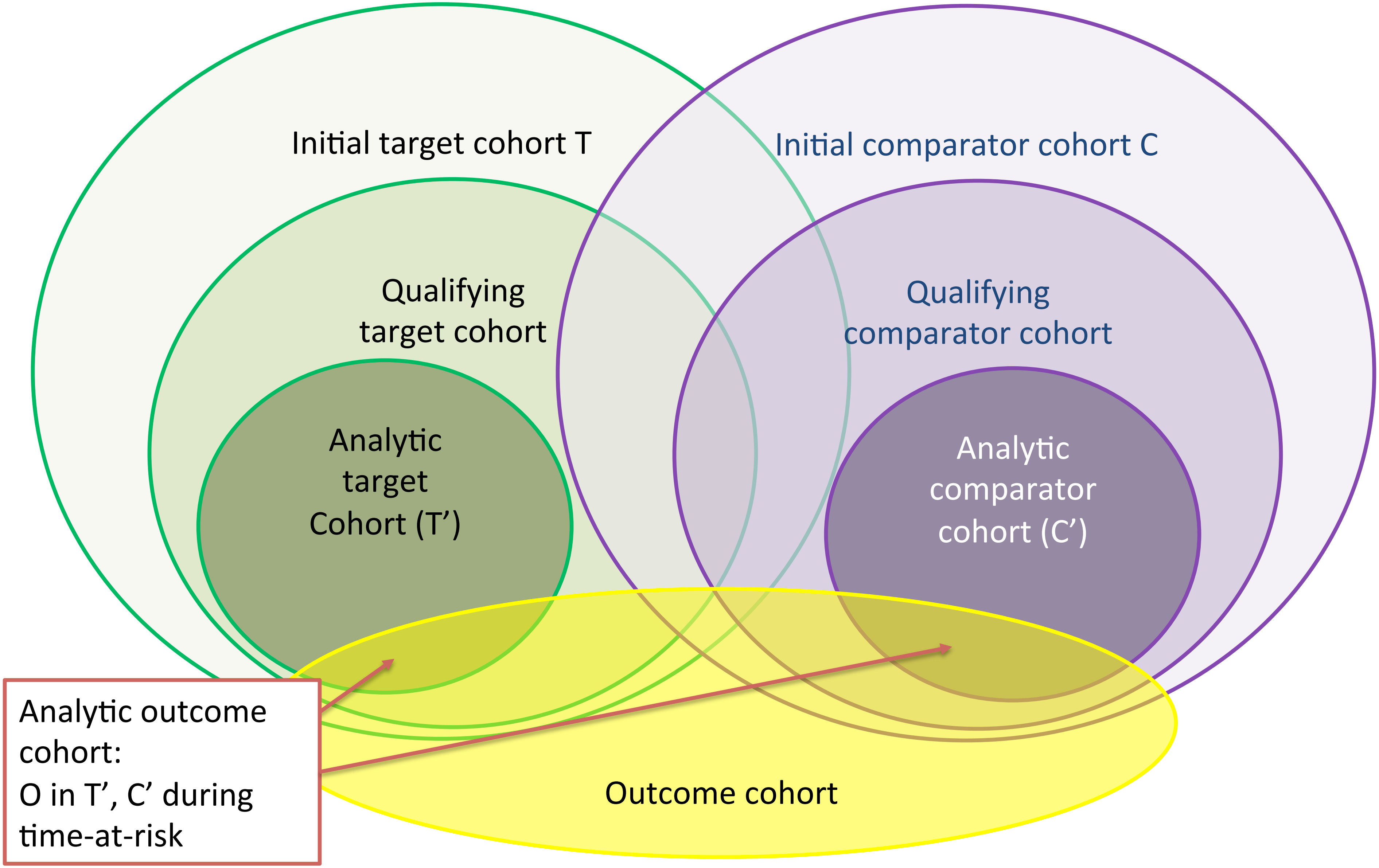
Observed

Unobserved

Unobserved

Observed

Cohort restriction in comparative cohort analyses


Initial target cohort T
Initial comparator cohort C
Qualifying target cohort
Qualifying comparator cohort
Analytic target Cohort (T')
Analytic comparator cohort (C')
Analytic outcome cohort: O in T', C' during time-at-risk
Outcome cohort

The choice of the outcome model defines your research question

	Logistic regression	Poisson regression	Cox proportional hazards
How the outcome cohort is used	Binary classifier of presence/ absence of outcome during the fixed time-at-risk period	Count the number of occurrences of outcomes during time-at-risk	Compute time-to-event from time-at-risk start until earliest of first occurrence of outcome or time-at-risk end, and track the censoring event (outcome or no outcome)
'Risk' metric	Odds ratio	Rate ratio	Hazard ratio
Key model assumptions	Constant probability in fixed window	Outcomes follow Poisson distribution with constant risk	Proportionality – constant relative hazard

When designing or reviewing a study, ask yourself:

Input parameter	Design choice
Target cohort (T)	
Comparator cohort (C)	
Outcome cohort (O)	
Time-at-risk	
Model specification	

Exercise 1

- Define your own problem

Break

Exercise 2

- Apply the framework to a published paper

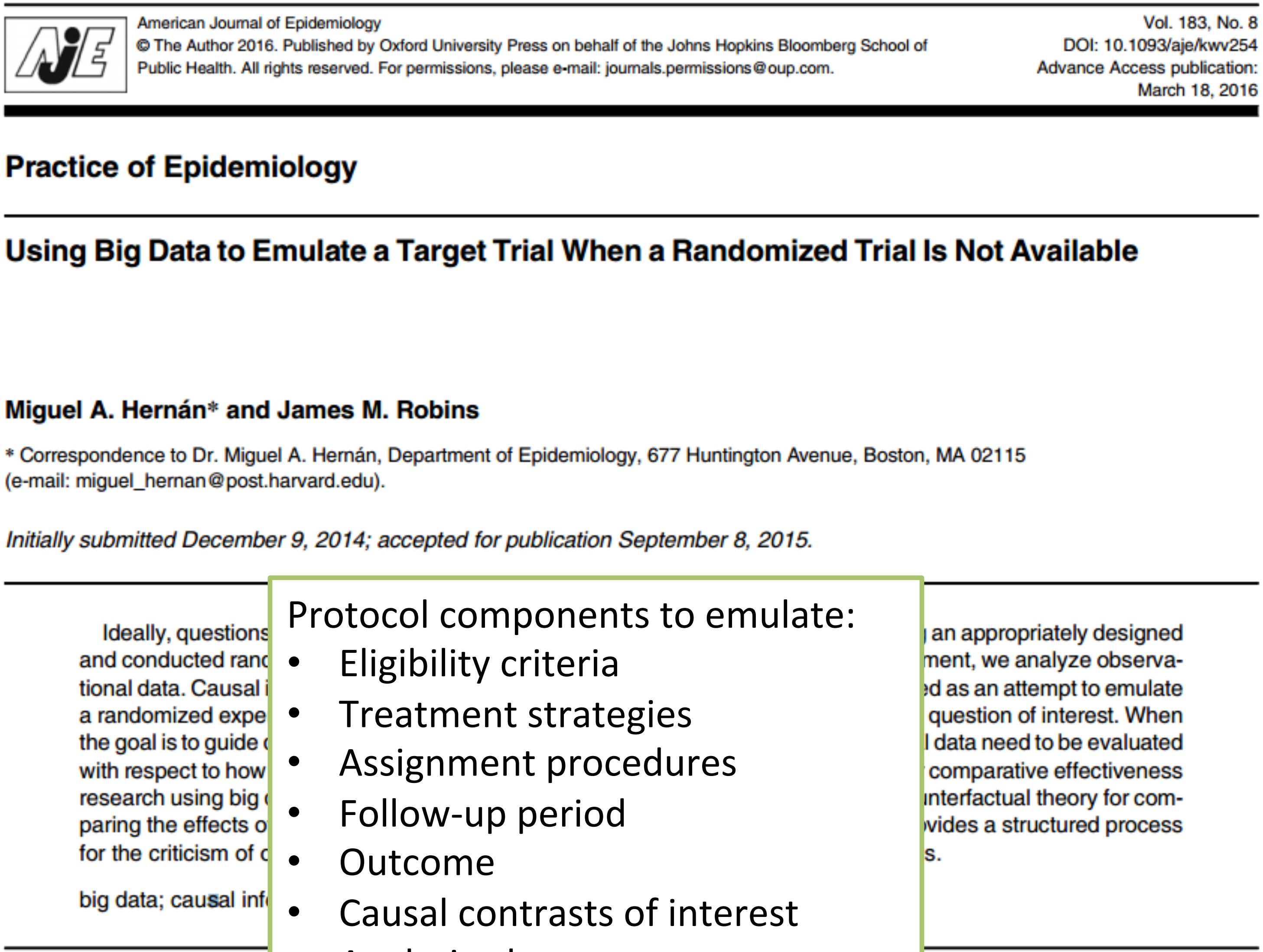
Observational study design Part #2

Patrick Ryan, PhD

Columbia University

Janssen Research and Development

Design an observational study like you would a randomized trial



American Journal of Epidemiology
© The Author 2016. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 183, No. 8
DOI: 10.1093/aje/kwv254
Advance Access publication:
March 18, 2016

Practice of Epidemiology

Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available

Miguel A. Hernán* and James M. Robins

* Correspondence to Dr. Miguel A. Hernán, Department of Epidemiology, 677 Huntington Avenue, Boston, MA 02115 (e-mail: miguel_hernan@post.harvard.edu).

Initially submitted December 9, 2014; accepted for publication September 8, 2015.

Ideally, questions an appropriately designed
and conducted ranc ment, we analyze observa-
tional data. Causal i ed as an attempt to emulate
a randomized expe question of interest. When
the goal is to guide I data need to be evaluated
with respect to how comparative effectiveness
research using big interfactual theory for com-
paring the effects o vides a structured process
for the criticism of c s.

big data; causal inf

Protocol components to emulate:

- Eligibility criteria
- Treatment strategies
- Assignment procedures
- Follow-up period
- Outcome
- Causal contrasts of interest
- Analysis plan

- Bias = expected value of the error distribution

$$\mathrm{BIAS}[\theta] = E[\theta - \theta] = E[\theta] - \theta$$

where θ = true value, θ = estimate of θ

- Mean squared error = metric to evaluate the quality of an estimator, accounting for both random and systematic error

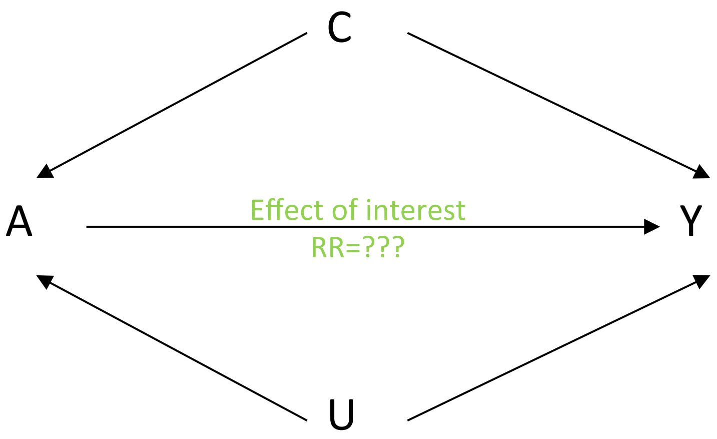
$$\mathrm{MSE}[\theta] = E[(\theta - \theta)\uparrow 2] = (\mathrm{BIAS}[\theta])\uparrow 2 + Var[\theta]$$

As studies increase in sample size, random error converges to 0 but systematic error still persists!

Types of systematic error

- Confounding
- Misclassification (Measurement error)
- Selection bias (generalizability)

Confounding



Challenge:
Producing an ‘unconfounded’ estimate relies on (empirically untestable) assumption that
1) all confounders were observable, and properly modeled in the design or analysis, and
2) no unobserved factors are associated with both exposure and outcome

A=exposure
Y=outcome

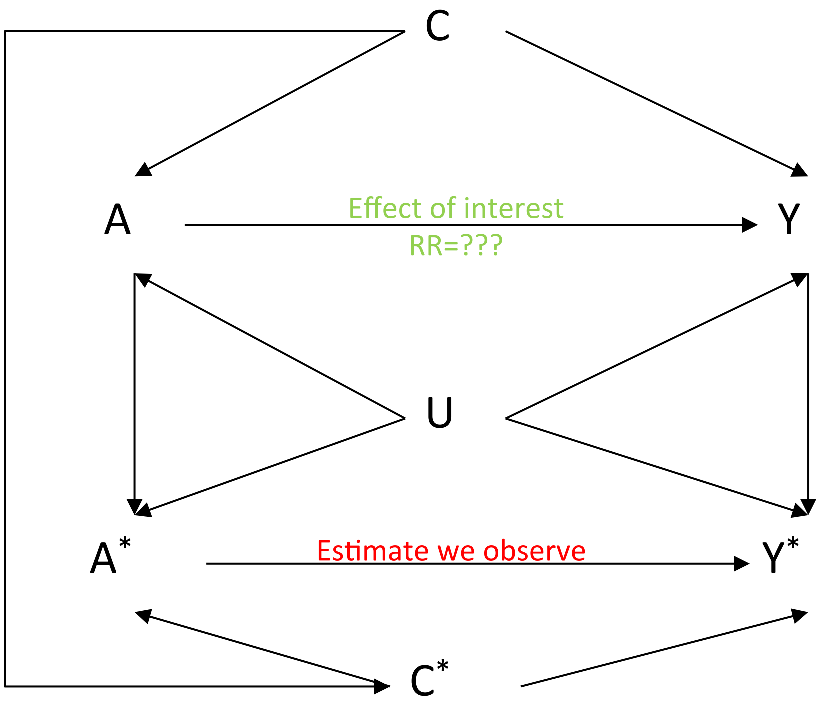
C = observed and modeled confounder

U = unobserved or mismodeled confounder

How do you assess confounding?

- PS distribution
- Covariate balance

Misclassification (measurement error)



A^*=proxy for exposure
Y^*=proxy for outcome
C^* = proxy for observed confounder

A=exposure
Y=outcome

C = observed and modeled confounder

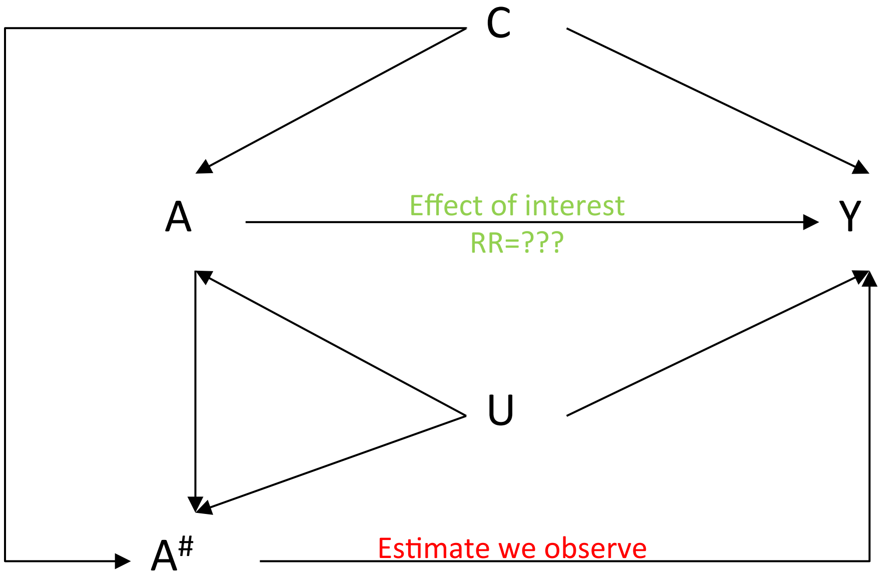
U = unobserved or mismodeled confounder

Challenge:
All observations are imperfect proxies for true patient status. Misclassification error can exist for all exposures, outcomes and covariates, but is generally unknown or not properly estimated (via sensitivity and specificity), and is rarely formally integrated into effect estimation.

How do you assess measurement error?

- Covariate summary for exposures
- Operating characteristics for outcome phenotype
  - Sensitivity
  - Specificity
  - Positive predictive value

Selection bias and generalizability



Challenge:
A database is a non-random sample of an underlying population. A cohort is a non-random sample of the database. Study design and analysis decisions may further restrict the cohort composition. Selection bias is rarely evaluated and often empirically untestable.

A# =non-random sample of exposure

A=exposure
Y=outcome

C = observed and modeled confounder

U = unobserved or mismodeled confounder

How do you assess selection bias?

- Attrition table
- Covariate summary (compare before to after)

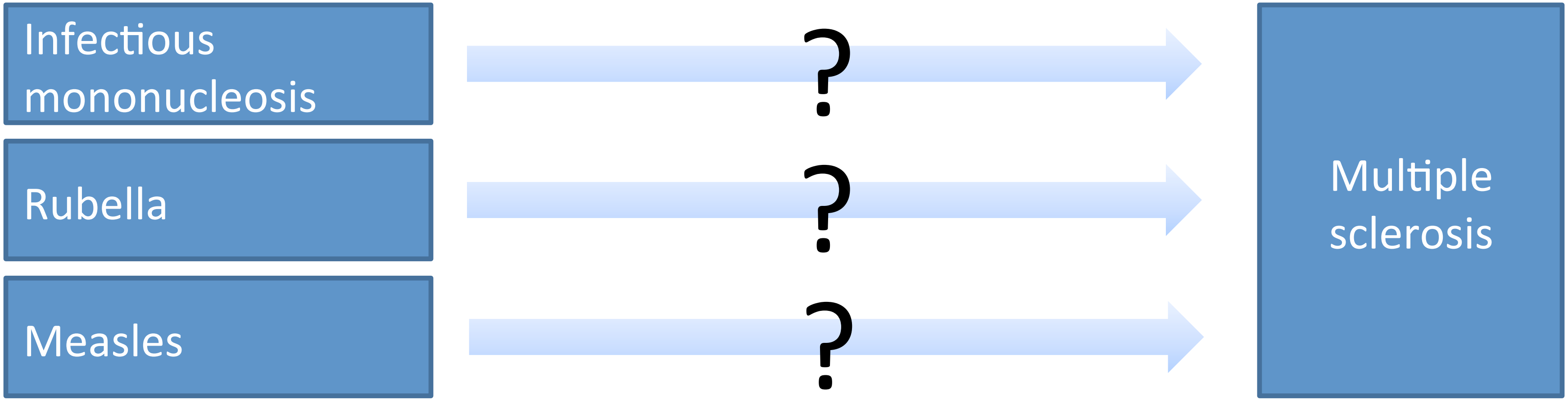
What can we do to address these challenges?

- Think really hard during study design and hope we get it right
- Equivocate in our summary of findings with a paragraph in the Discussion that reads:
  - "This study has several limitations. First, since this study relied on claims data, we had no data on <unobserved confounders>. Second, while we adjusted for <observed confounders>, residual confounding cannot be ruled out. Third, there is a potential for outcome misclassification… Fourth, there is a potential for duplicate person-years between <databases>. Lastly, as the mean follow-up was <short>, long-term effects may need to be further examined." (Kim et al., Arthritis & Rheumatology, 2017)
-
-

What can we do to address these challenges?

- Think really hard during study design and hope we get it right
- Equivocate in our summary of findings with a paragraph in the Discussion that reads:
  - "This study has several limitations. First, since this study relied on claims data, we had no data on <unobserved confounders>. Second, while we adjusted for <observed confounders>, residual confounding cannot be ruled out. Third, there is a potential for outcome misclassification… Fourth, there is a potential for duplicate person-years between <databases>. Lastly, as the mean follow-up was <short>, long-term effects may need to be further examined." (Kim et al., Arthritis & Rheumatology, 2017)
- Perform diagnostic analyses that attempt to detect if residual error may still be present
- Quantify magnitude of residual error and calibrate statistics

Examples of negative controls



Infectious mononucleosis

Rubella

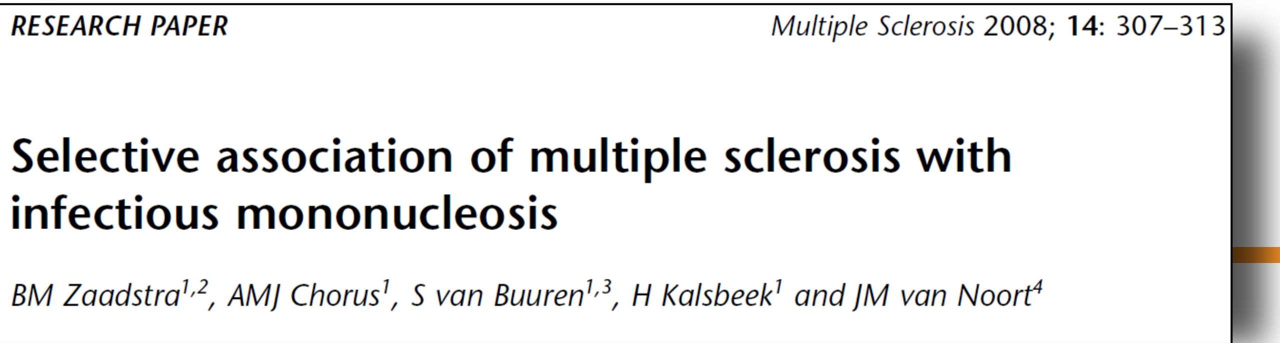
Measles

?

?

?

Multiple sclerosis



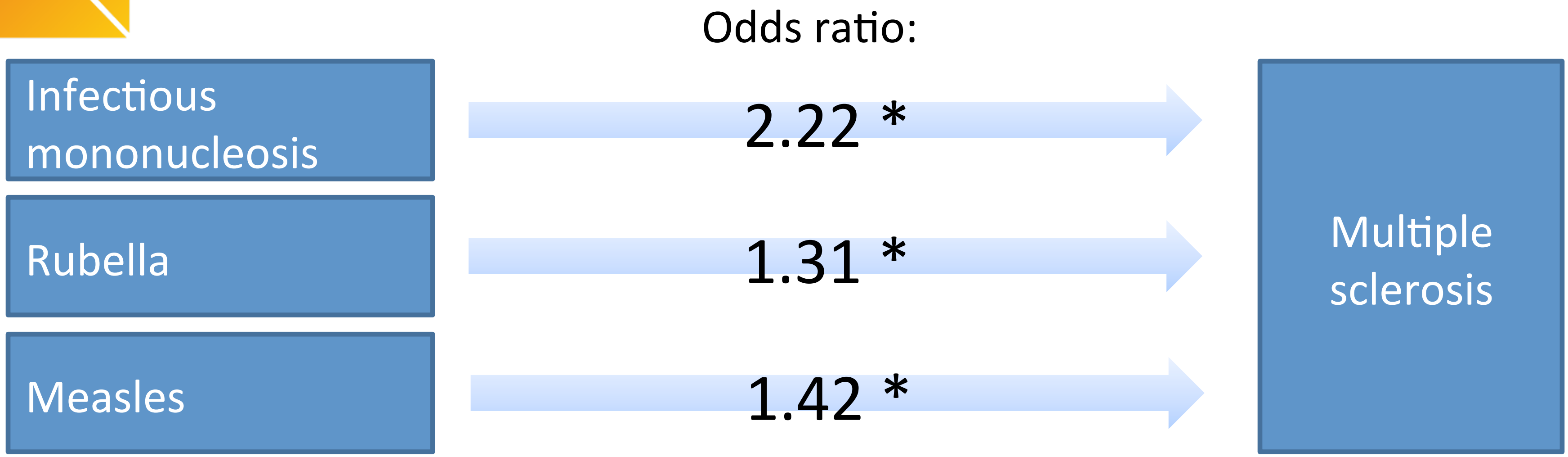
RESEARCH PAPER

Multiple Sclerosis 2008; **14**: 307–313

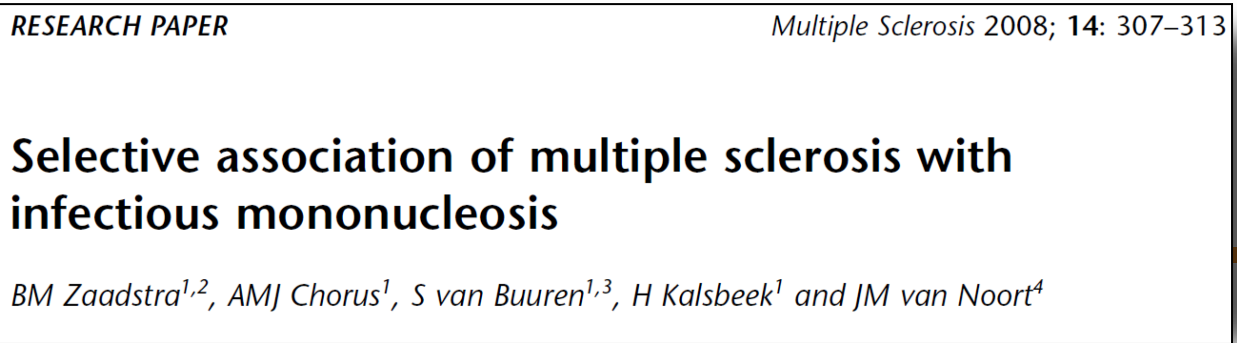
Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control



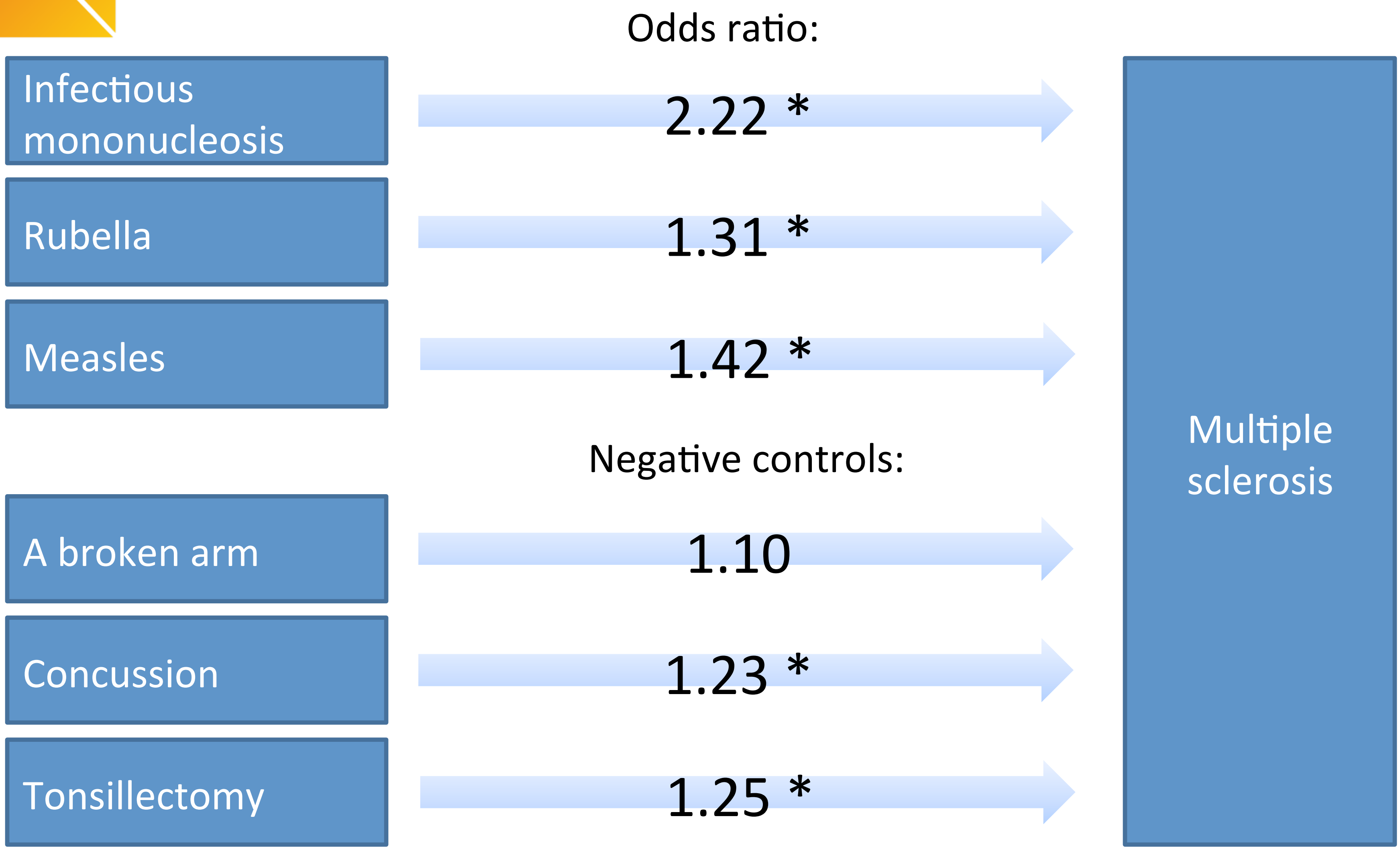
* P < .05


RESEARCH PAPER *Multiple Sclerosis* 2008; **14**: 307–313

Selective association of multiple sclerosis with infectious mononucleosis

BM Zaadstra[1,2], AMJ Chorus[1], S van Buuren[1,3], H Kalsbeek[1] and JM van Noort[4]

Example of a negative control



Odds ratio:

Infectious mononucleosis → 2.22 *

Rubella → 1.31 *

Measles → 1.42 *

Negative controls:

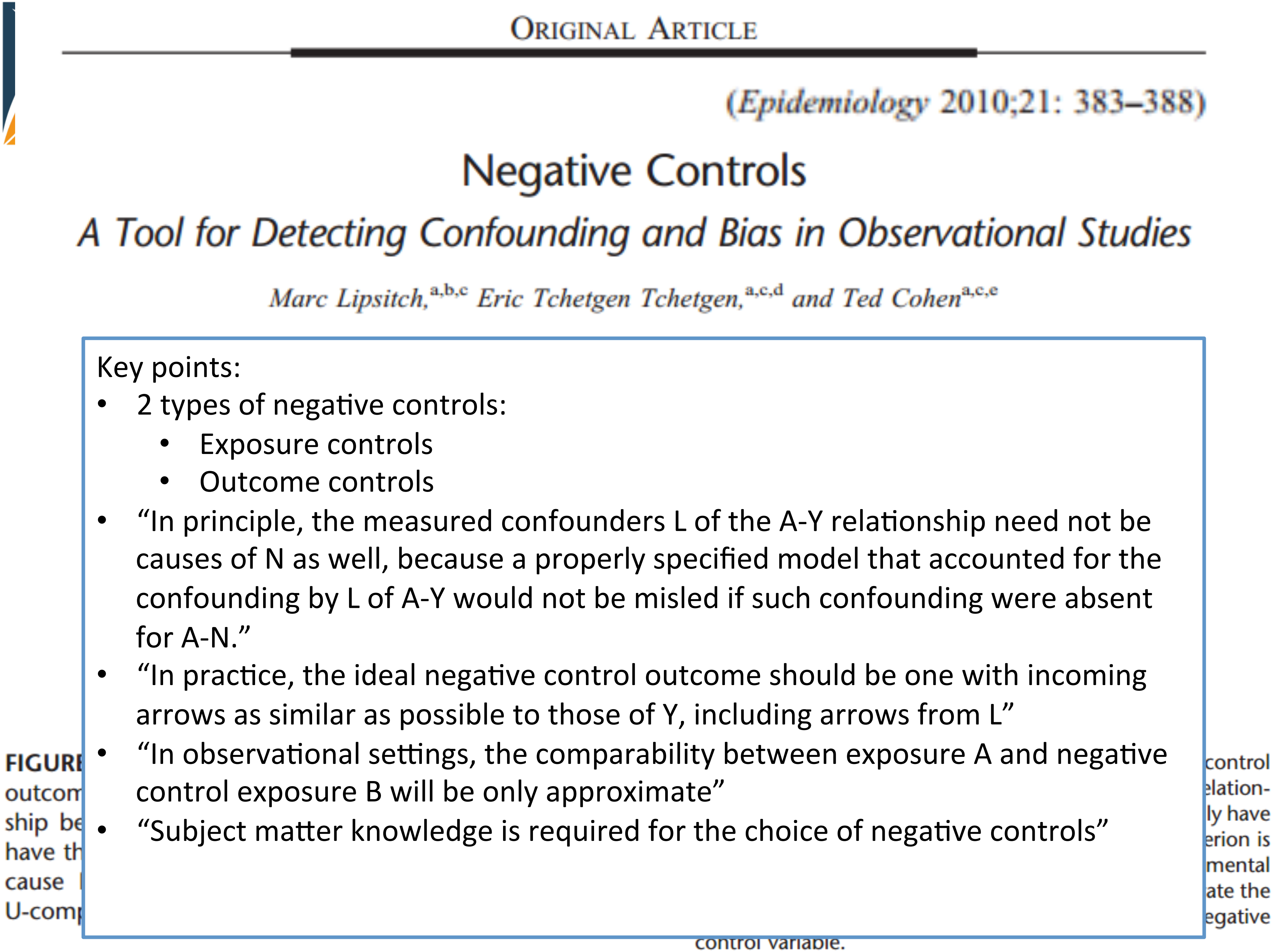
A broken arm → 1.10

Concussion → 1.23 *

Tonsillectomy → 1.25 *

Multiple sclerosis

* P < .05



ORIGINAL ARTICLE

(*Epidemiology* 2010;21: 383–388)

Negative Controls

A Tool for Detecting Confounding and Bias in Observational Studies

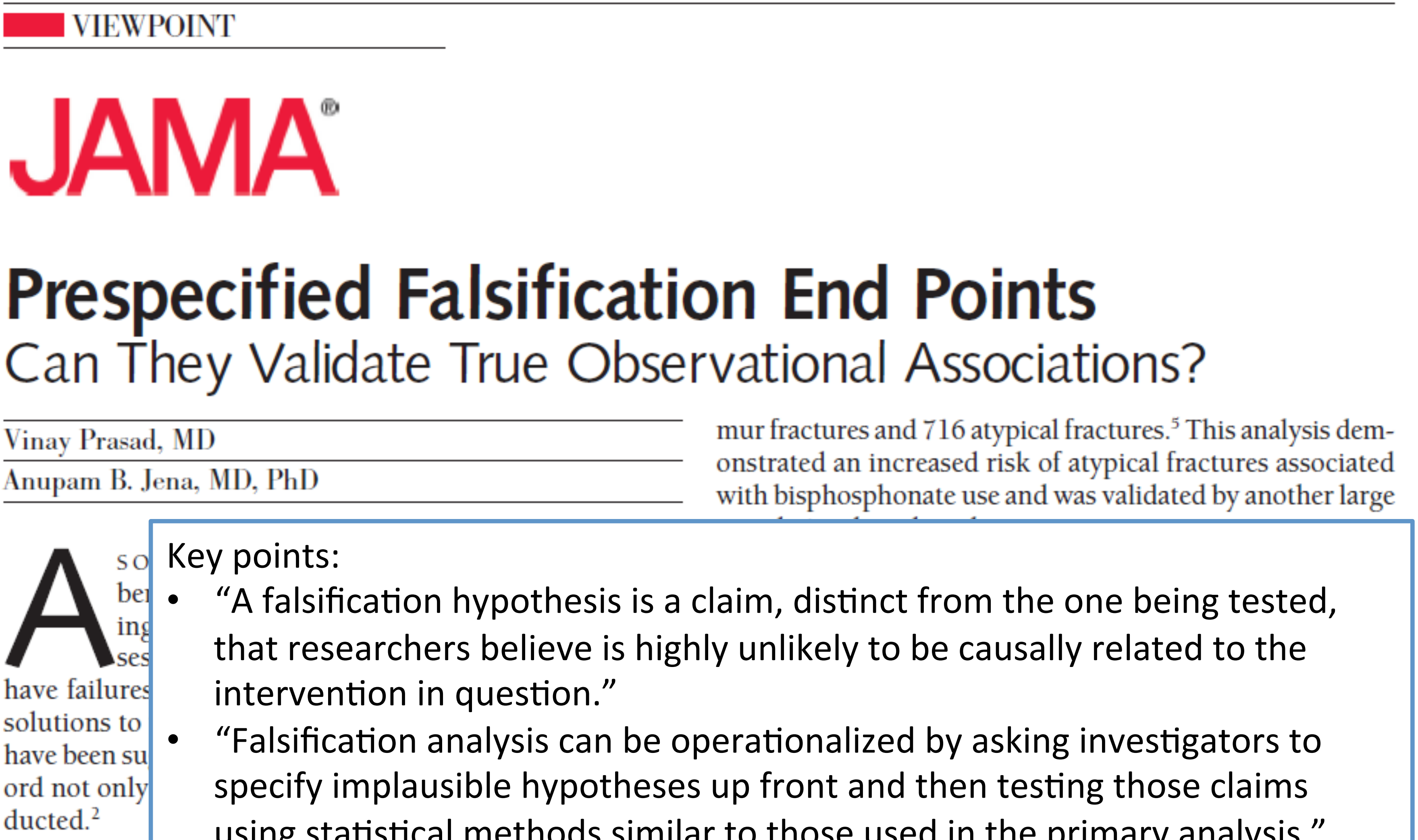
Marc Lipsitch,[a,b,c] *Eric Tchetgen Tchetgen,*[a,c,d] *and Ted Cohen*[a,c,e]

Key points:

- 2 types of negative controls:
  - Exposure controls
  - Outcome controls
- "In principle, the measured confounders L of the A-Y relationship need not be causes of N as well, because a properly specified model that accounted for the confounding by L of A-Y would not be misled if such confounding were absent for A-N."
- "In practice, the ideal negative control outcome should be one with incoming arrows as similar as possible to those of Y, including arrows from L"
- "In observational settings, the comparability between exposure A and negative control exposure B will be only approximate"
- "Subject matter knowledge is required for the choice of negative controls"

FIGURE
outcom
ship be
have th
cause
U-comp

control
elation-
ly have
erion is
mental
ate the
egative
control variable.



VIEWPOINT

JAMA®

Prespecified Falsification End Points

Can They Validate True Observational Associations?

Vinay Prasad, MD

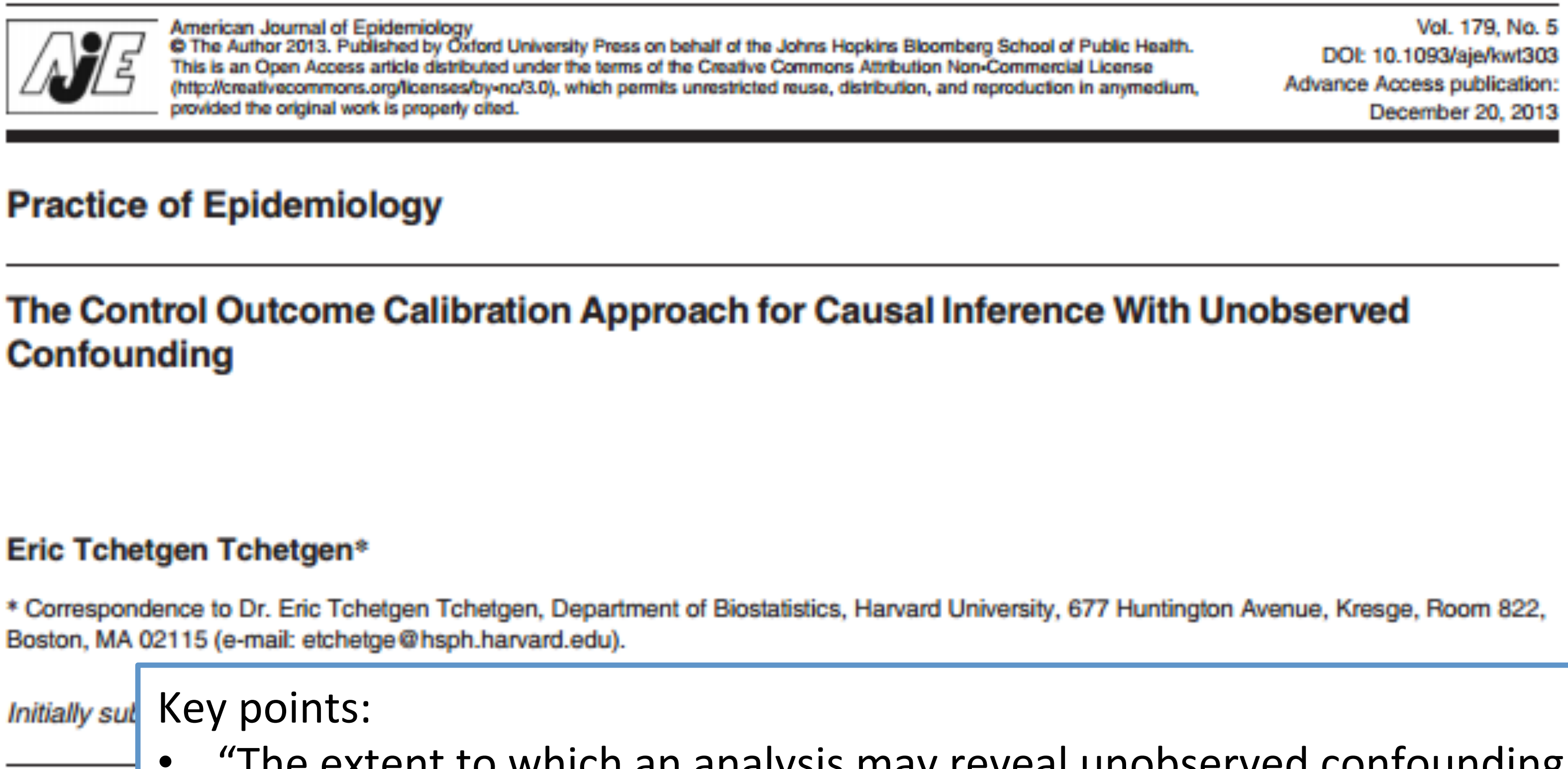
Anupam B. Jena, MD, PhD

mur fractures and 716 atypical fractures.[5] This analysis demonstrated an increased risk of atypical fractures associated with bisphosphonate use and was validated by another large

A so
bei
ing
ses
have failures
solutions to
have been su
ord not only
ducted.[2]

Key points:

- “A falsification hypothesis is a claim, distinct from the one being tested, that researchers believe is highly unlikely to be causally related to the intervention in question.”
- “Falsification analysis can be operationalized by asking investigators to specify implausible hypotheses up front and then testing those claims using statistical methods similar to those used in the primary analysis.”
- “Although no published recommendations exist, standardized falsification analyses with 3 or 4 prespecified or highly prevalent disease outcomes may help strengthen the validity of observational studies”



American Journal of Epidemiology

Vol. 179, No. 5
DOI: 10.1093/aje/kwt303
Advance Access publication:
December 20, 2013

Practice of Epidemiology

The Control Outcome Calibration Approach for Causal Inference With Unobserved Confounding

Eric Tchetgen Tchetgen*

* Correspondence to Dr. Eric Tchetgen Tchetgen, Department of Biostatistics, Harvard University, 677 Huntington Avenue, Kresge, Room 822, Boston, MA 02115 (e-mail: etchetge@hsph.harvard.edu).

Initially sub

Key points:

- "The extent to which an analysis may reveal unobserved confounding bias relies on the non-empirically verifiable assumption that the negative control outcome is carefully chosen so that it is solely influenced by observed and unobserved confounders of the exposure-outcome relationship in view"
- "We propose to use a negative control outcome not only to detect, but also to correct for unmeasured confounding bias"

OPEN

(*Epidemiology* 2016;27:637–641)

Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies

Benjamin F. Arnold, Ayse Ercumen, Jade Benjamin-Chung, and John M. Colford, Jr

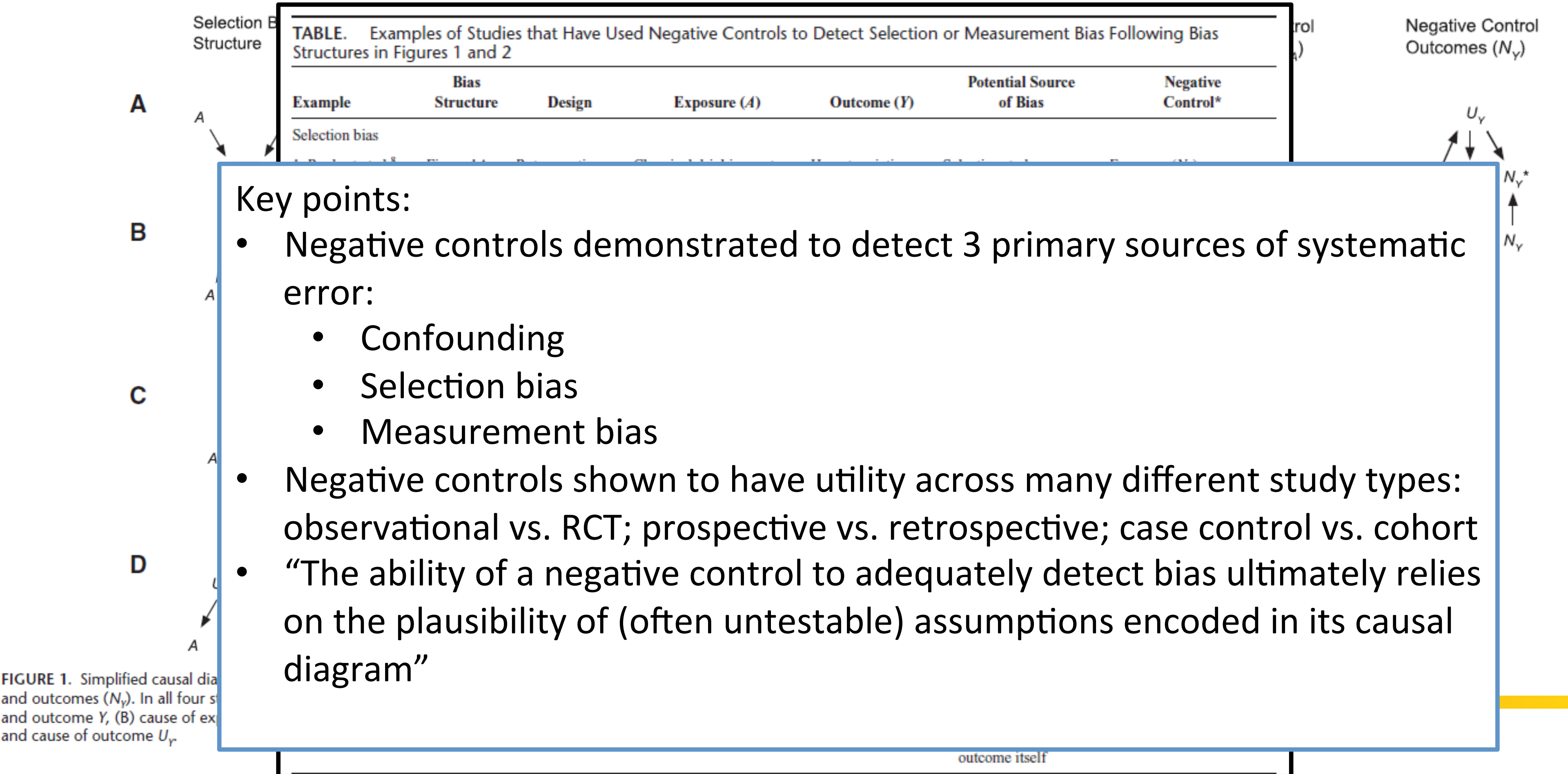


TABLE. Examples of Studies that Have Used Negative Controls to Detect Selection or Measurement Bias Following Bias Structures in Figures 1 and 2

Example	Bias Structure	Design	Exposure (A)	Outcome (Y)	Potential Source of Bias	Negative Control*
Selection bias						
					outcome itself	

FIGURE 1. Simplified causal dia
and outcomes (N_Y). In all four s
and outcome Y, (B) cause of ex
and cause of outcome U_Y.

Key points:

- Negative controls demonstrated to detect 3 primary sources of systematic error:
  - Confounding
  - Selection bias
  - Measurement bias
- Negative controls shown to have utility across many different study types: observational vs. RCT; prospective vs. retrospective; case control vs. cohort
- “The ability of a negative control to adequately detect bias ultimately relies on the plausibility of (often untestable) assumptions encoded in its causal diagram”

Special Issue Paper

Statistics in Medicine

Received 4 November 2011, Accepted 28 August 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5620

Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership[‡]

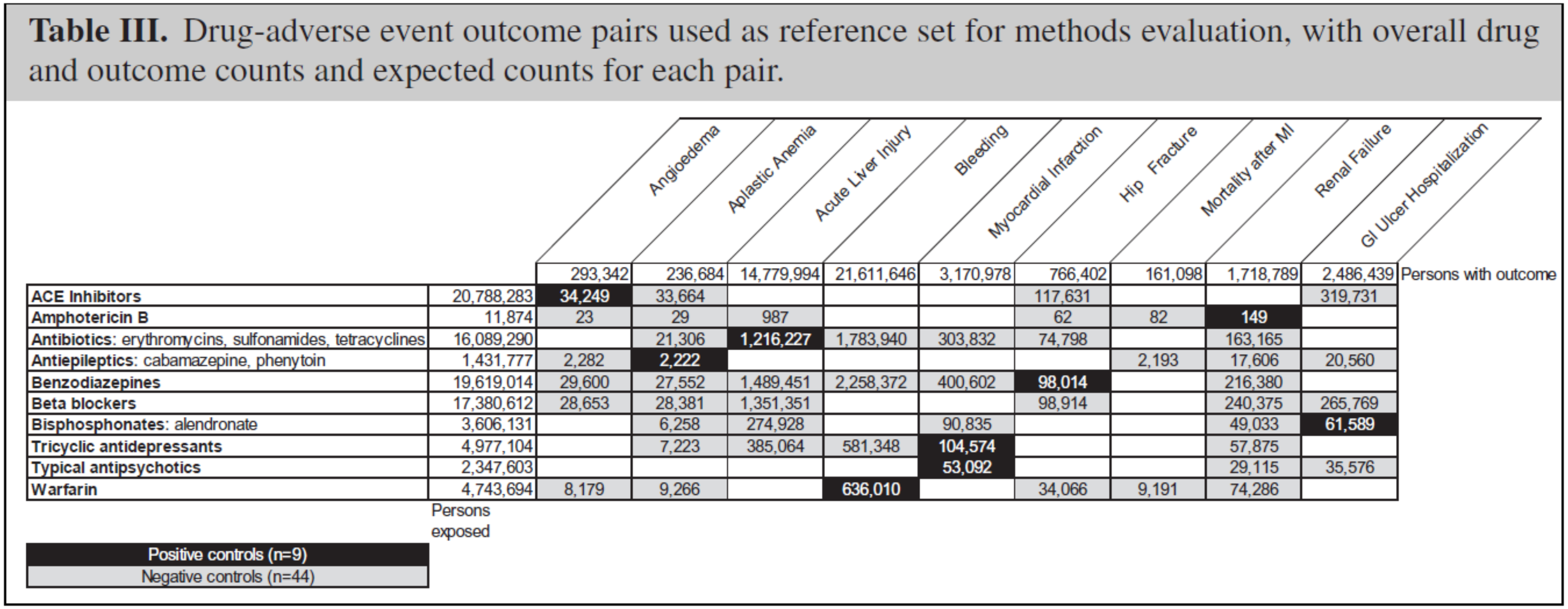


Table III. Drug-adverse event outcome pairs used as reference set for methods evaluation, with overall drug and outcome counts and expected counts for each pair.

		Angioedema	Aplastic Anemia	Acute Liver Injury	Bleeding	Myocardial Infarction	Hip Fracture	Mortality after MI	Renal Failure	GI Ulcer Hospitalization	
		293,342	236,684	14,779,994	21,611,646	3,170,978	766,402	161,098	1,718,789	2,486,439	Persons with outcome
ACE Inhibitors	20,788,283	■ 34,249	33,664				117,631			319,731	
Amphotericin B	11,874	23	29	987			62	82	■ 149		
Antibiotics: erythromycins, sulfonamides, tetracyclines	16,089,290		21,306	■ 1,216,227	1,783,940	303,832	74,798		163,165		
Antiepileptics: cabamazepine, phenytoin	1,431,777	2,282	■ 2,222					2,193	17,606	20,560	
Benzodiazepines	19,619,014	29,600	27,552	1,489,451	2,258,372	400,602	■ 98,014		216,380		
Beta blockers	17,380,612	28,653	28,381	1,351,351			98,914		240,375	265,769	
Bisphosphonates: alendronate	3,606,131		6,258	274,928		90,835			49,033	■ 61,589	
Tricyclic antidepressants	4,977,104		7,223	385,064	581,348	■ 104,574			57,875		
Typical antipsychotics	2,347,603					■ 53,092			29,115	35,576	
Warfarin	4,743,694	8,179	9,266		■ 636,010		34,066	9,191	74,286		
	Persons exposed										

■ Positive controls (n=9)

Negative controls (n=44)

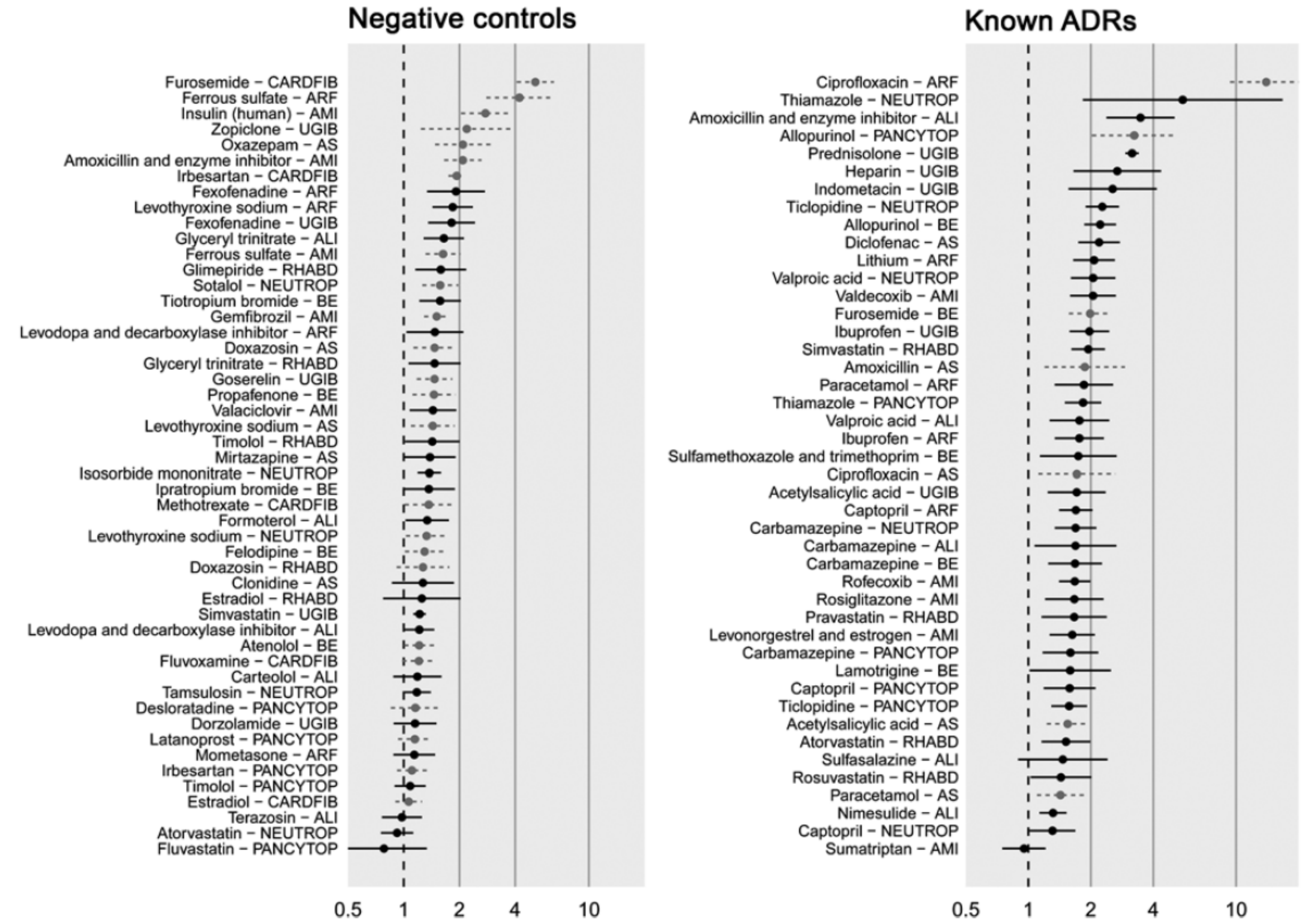
ORIGINAL ARTICLE

Medical Care • Volume 50, Number 10, October 2012

Using Electronic Health Care Records for Drug Safety Signal Detection

A Comparative Evaluation of Statistical Methods

R
David
Loren

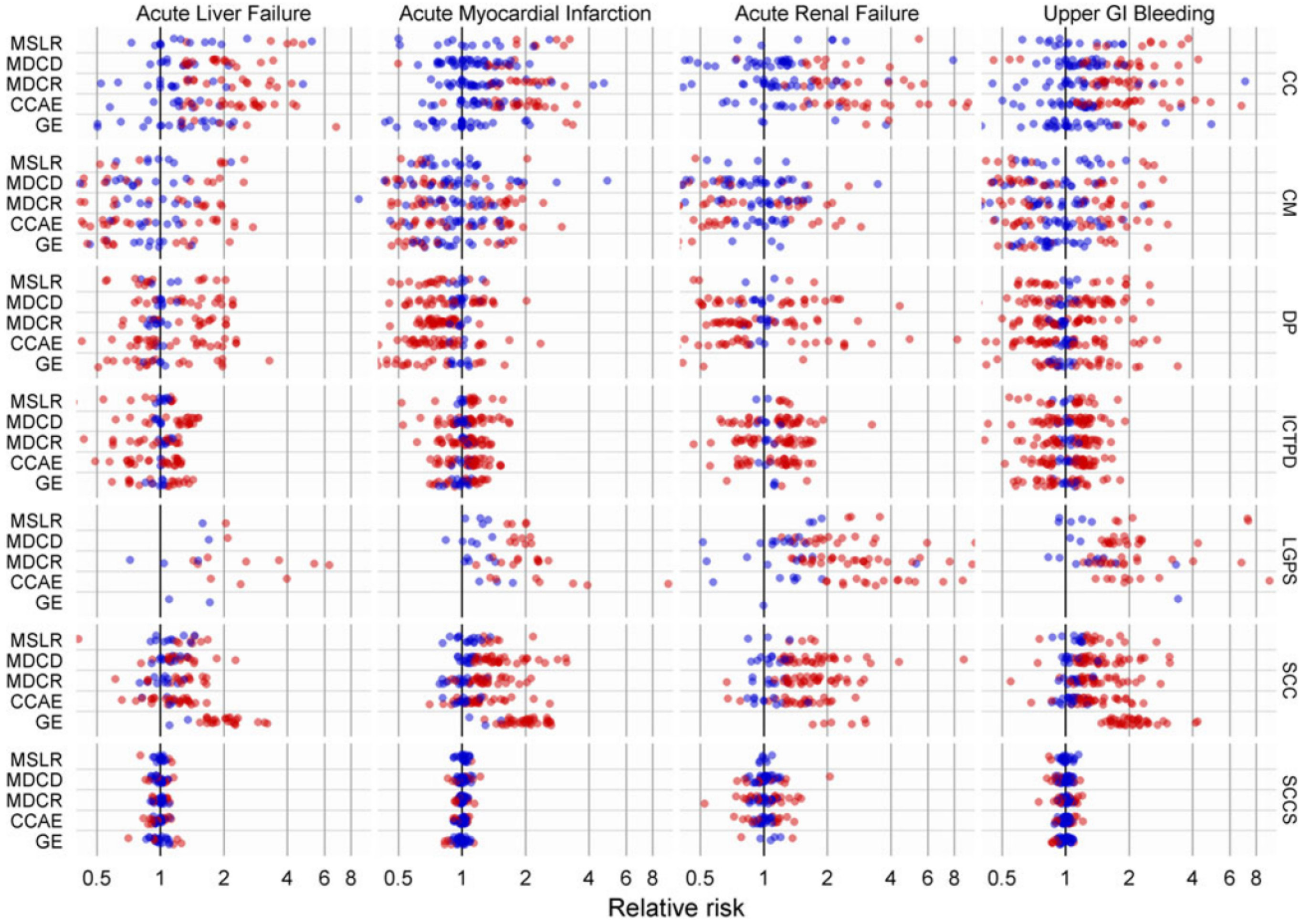


#
)*

Drug Saf (2013) 36 (Suppl 1):S143–S158
DOI 10.1007/s40264-013-0108-9

ORIGINAL RESEARCH ARTICLE

A Comparison of the Empirical Performance of Methods for a Risk Identification System



Research Article

Statistics in Medicine

Received 12 November 2012, Accepted 3 July 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5925

Interpreting observational studies: why empirical calibration is needed to correct *p*-values

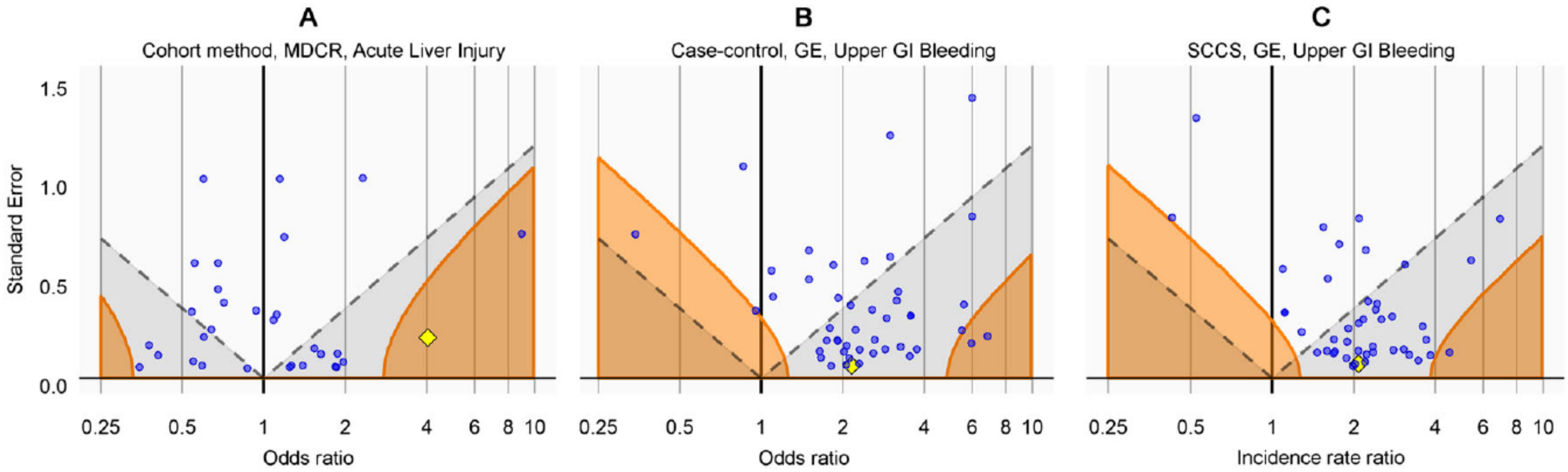
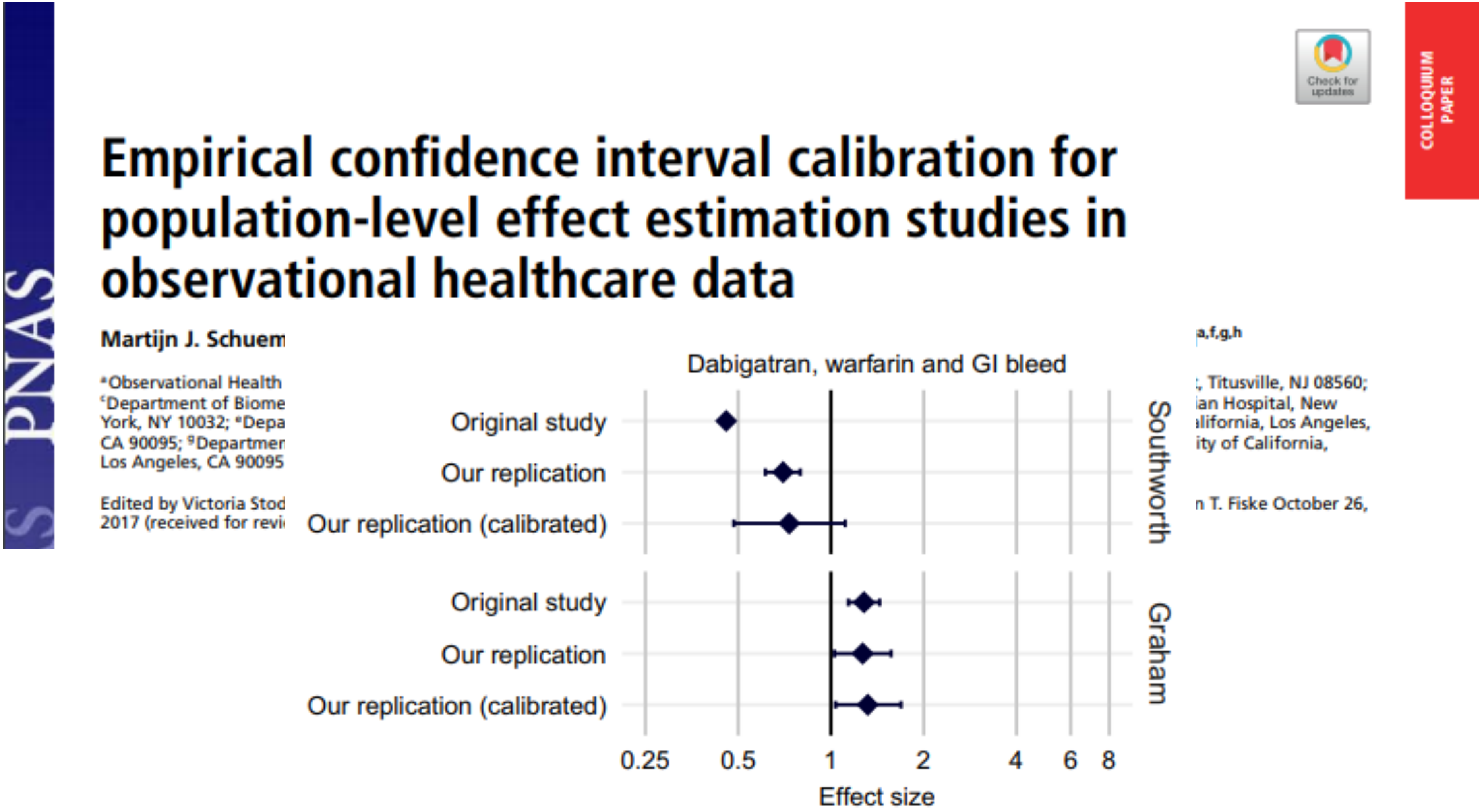


Figure 3. Traditional and calibrated significance testing. Estimates below the dashed line (gray area) have $p < 0.05$ using traditional p-value calculation. Estimates in the orange areas have $p < 0.05$ using the calibrated p-value calculation. Blue dots indicate negative controls, and the yellow diamond indicates the drugs of interest: isoniazid (A) and sertraline (B and C).



COLLOQUIUM PAPER

Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data

Martijn J. Schuem a,f,g,h

ªObservational Health , Titusville, NJ 08560;
ᶜDepartment of Biome ian Hospital, New
York, NY 10032; ᵉDepa alifornia, Los Angeles,
CA 90095; ᵍDepartmer ity of California,
Los Angeles, CA 90095

Edited by Victoria Stod n T. Fiske October 26,
2017 (received for revi

Fig. 5. Estimates from the original studies and our reproduction of the studies by Southworth et al. (12) and Graham et al. (13) both before and after calibration.

Exercise 3

- Evaluate Graham, what did they do to mitigate the threat of systematic error? How do you know they were successful?