The journey toward Population-level Effect Estimation

Martijn Schuemie, PhD
Janssen Research and Development

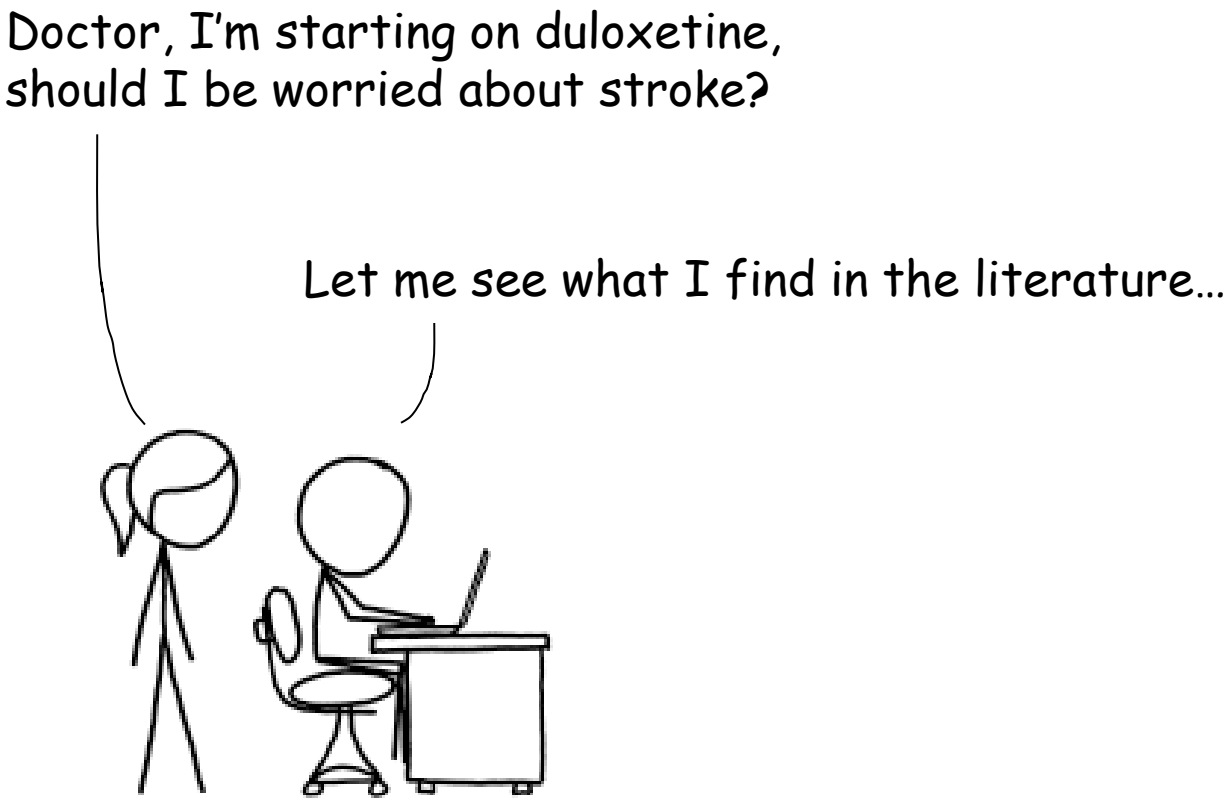
Population-level effect estimation

- What is the effect of treatment A on outcome X?

- What is the effect of treatment A on outcome X, compared to exposure B?

Population-level effect estimation

- How to produce evidence from the data?
- How do we know the evidence is reliable?
- How do we share evidence to inform decision making?


Doctor, I'm starting on duloxetine, should I be worried about stroke?
Let me see what I find in the literature…

Evidence from literature

Paper by Lee et al, 2016

- Compare new users of SNRIs (includes duloxetine) vs SSRIs
- Taiwanese insurance claims data
- 12 month washout
- remove people using both drugs
- remove people with a prior history of head injury
- remove people with a prior history of stroke or intracranial hemorrhage
- Propensity score: logistic regression with treatment as dependent variable
- HOI is Stroke: first hospitalization with ICD-9 433,434, or 436
- time-varying Cox regression using 5 PS strata

	Crude Hazard Ratio (95% CI)	*P*	Adjusted Hazard Ratio[a] (95% CI)	*P*
Main analyses				
SNRIs (n = 76,920) vs SSRIs (n = 582,650)				
Ischemic stroke	0.92 (0.83–1.02)	.12	1.01 (0.90–1.12)	.91

Lu, MSc[c]; Chia-Hsuin Chang, MD, ScD[c,d,*]; and Jou-Wei Lin, MD, PhD[e]

How reliable is this evidence?

- Can the results be reproduced?
- Did the analysis program do what it was supposed to do?
- Is the estimate unbiased?
- Does the p-value have nominal characteristics?
- Does the confidence interval really represent the uncertainty about the effect size?

  Are we really 95% confident the true effect size is between 0.90 and 1.12?

Population-level effect estimation

- How to produce evidence from the data?

‘Replicating’ Lee et al.

Our replication:

- Compare new users of Duloxetine (SNRI) vs. Sertraline (SSRI)
- US insurance claims data (Truven CCAE)
- 12 month washout
- remove people using both drugs
- remove people with a prior history of stroke
- restricted to people with a diagnosis of major depressive disorder and no prior diagnosis of bipolar disorder or schizophrenia
- Propensity score: regularized logistic regression with treatment as dependent variable, and used 58,285 covariates
- HOI is Stroke: first hospitalization with ICD-9 433,434, or 436 (but then coded as standard concepts)
- fixed-time Cox regression using 10 PS strata

OHDSI recommendations for evidence generation

- ✓ Post protocol online
  - Prespecify research objectives and design decisions
- ✓ Make study code open source
  - From CDM to hazard ratios
- ✓ Use validated software
  - OHDSI Methods Library uses unit tests and simulation
- ✓ Replicate across several databases
  - 4 included so far, more will follow

https://github.com/OHDSI/StudyProtocols/LargeScalePopEst

Population-level effect estimation

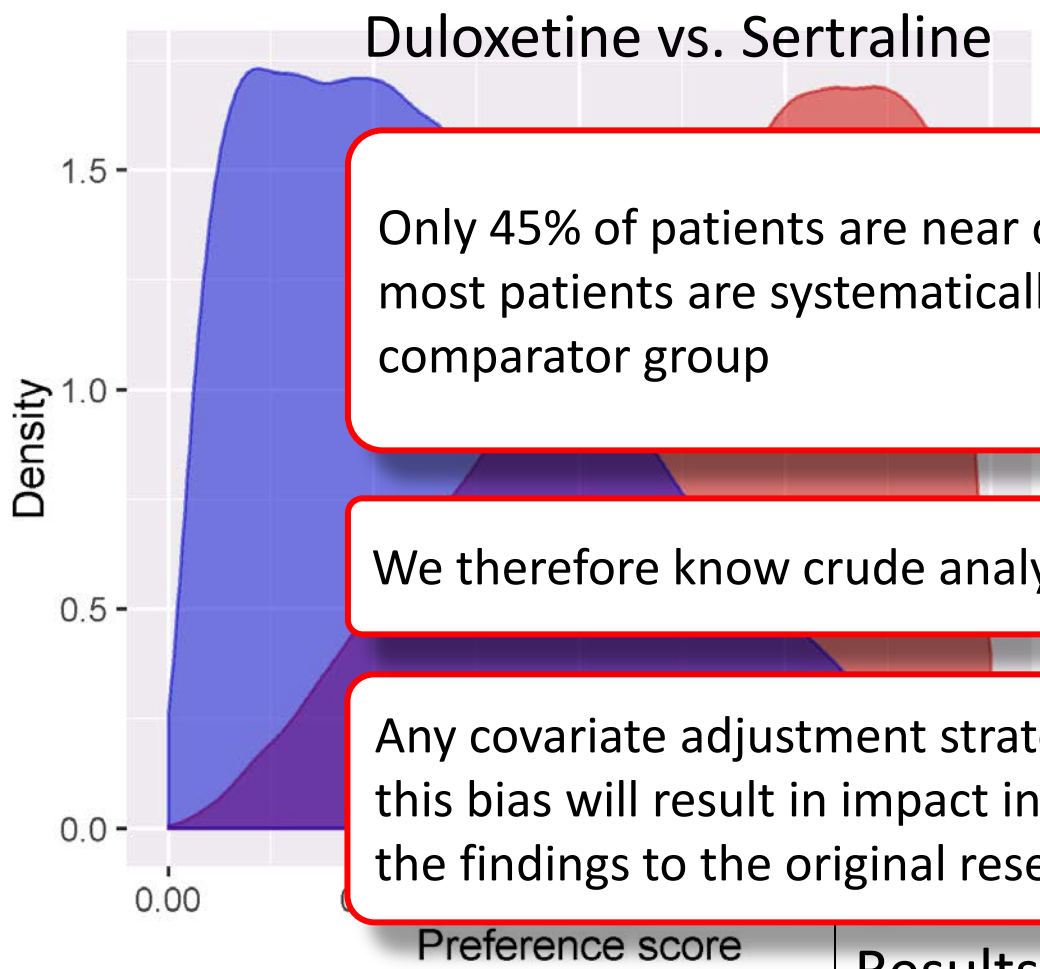
- How do we know the evidence is reliable?

Standard diagnostics

Most study designs have diagnostics that could be used, e.g.

- Propensity score distribution overlap
- Covariate balance

Diagnose the propensity score distribution



Only 45% of patients are near clinical equipoise, most patients are systematically different from the comparator group

We therefore know crude analysis will likely be biased

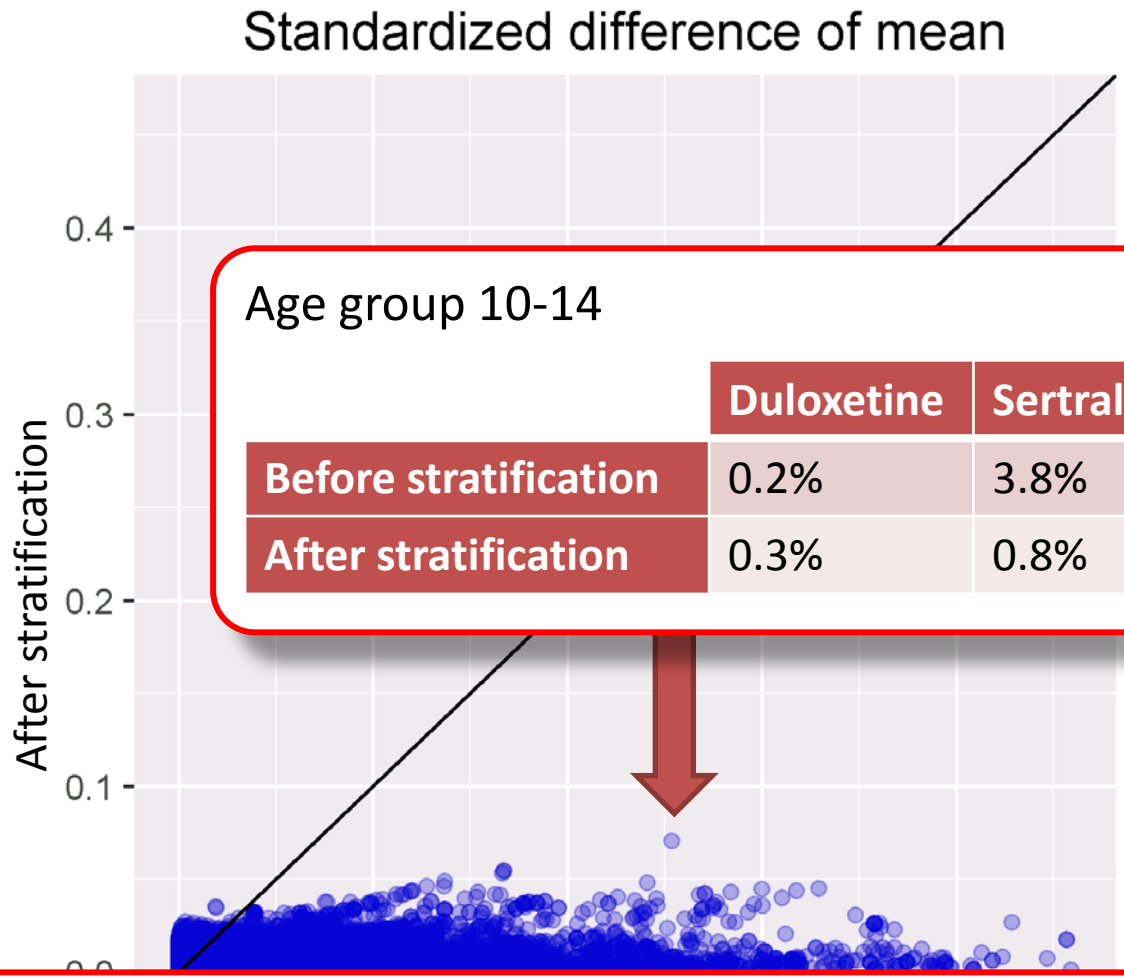
Any covariate adjustment strategy that corrects for this bias will result in impact in the generalizability of the findings to the original research question

Results from Truven CCAE

Duloxetine: n = 90,043

Sertraline: n = 175,950

Diagnose covariate balance



Age group 10-14

	Duloxetine	Sertraline
Before stratification	0.2%	3.8%
After stratification	0.3%	0.8%

After stratification on the propensity score, all 58,285 covariates have standardized difference of mean < 0.1

Empirical evaluation of the study

- Control

  exposure-outcome for which the effect size is known

- Negative control

  exposure-outcome where relative risk is believed to be 1

- Negative controls for comparative effectiveness

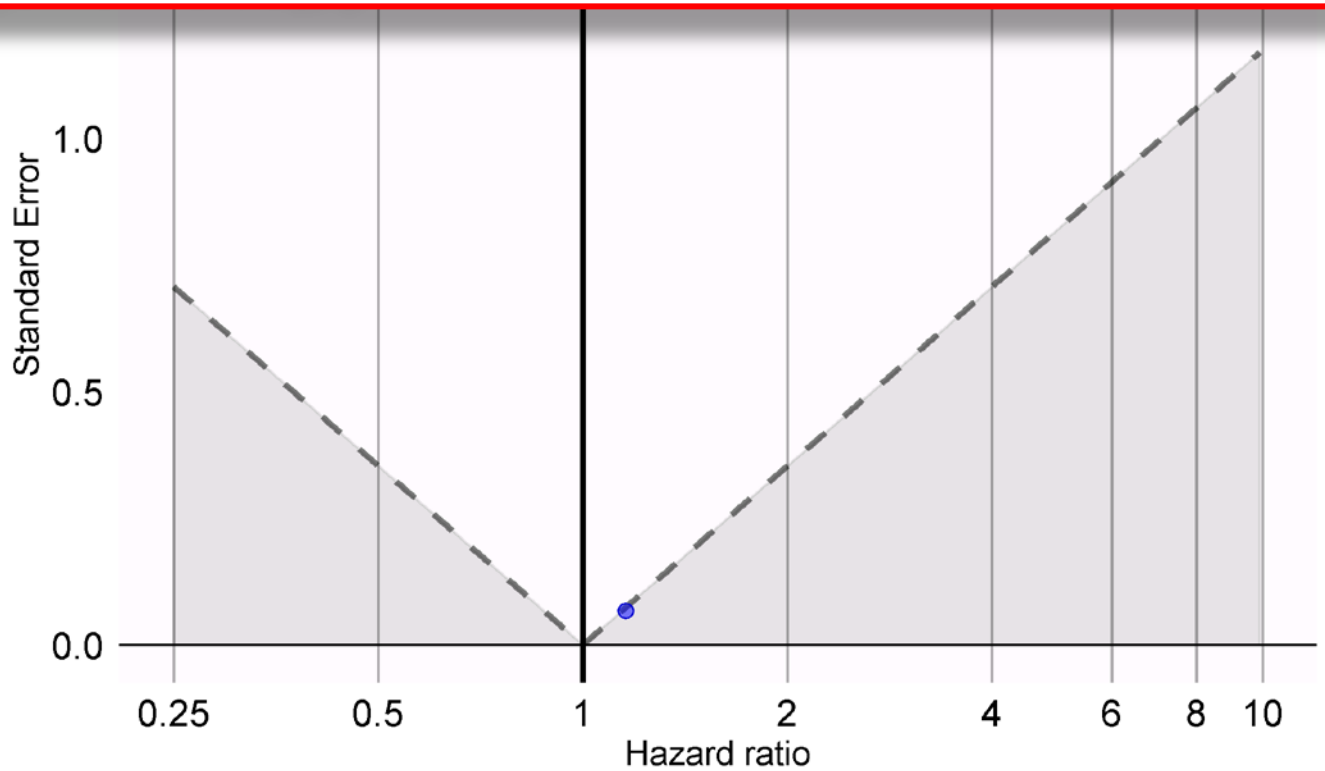
  outcomes not believed to be caused by either treatments

Example: ingrowing nail

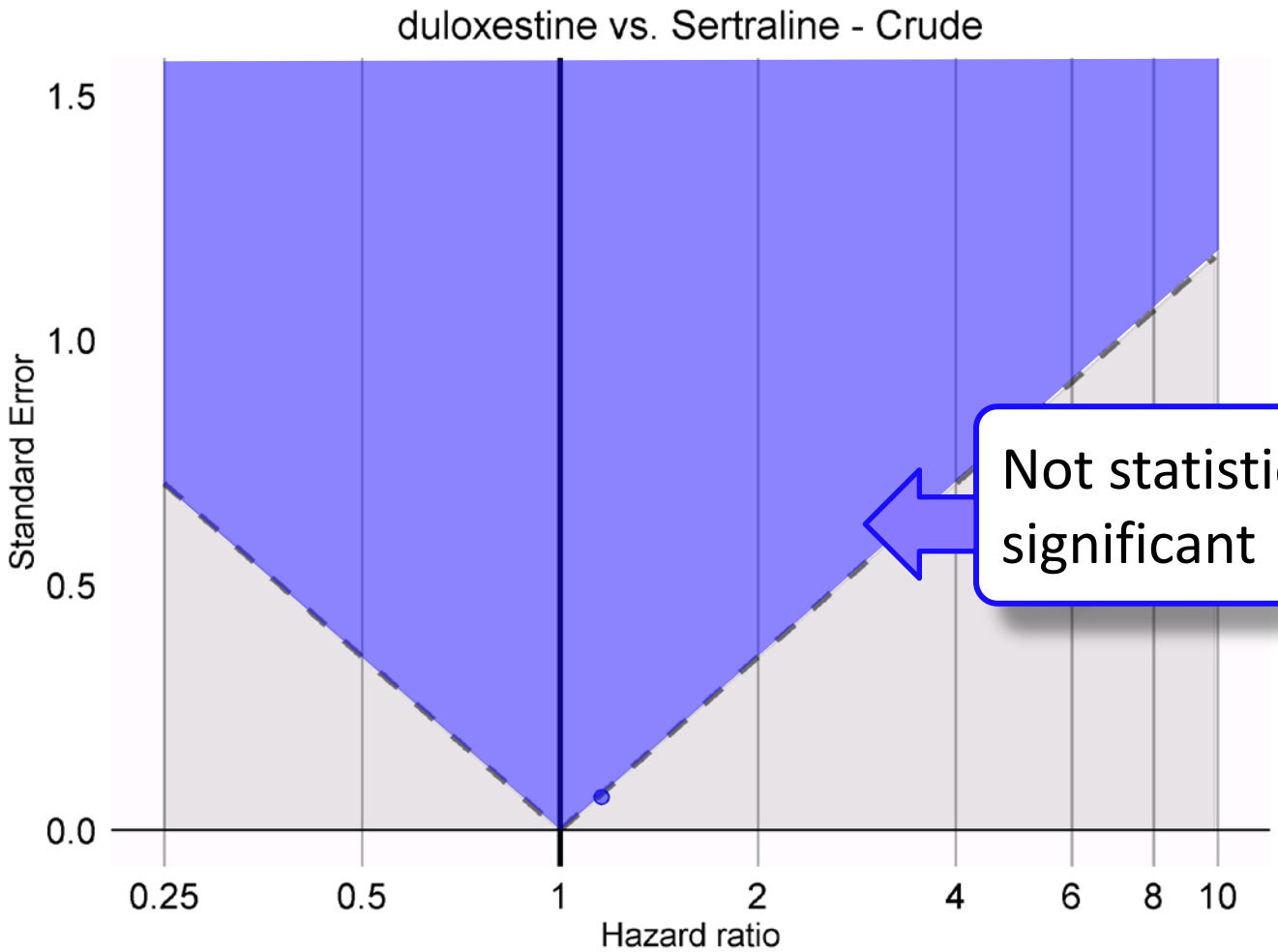
Negative control: ingrowing nail

Crude estimate:

HR = 1.16 (1.01 – 1.32), p = 0.03

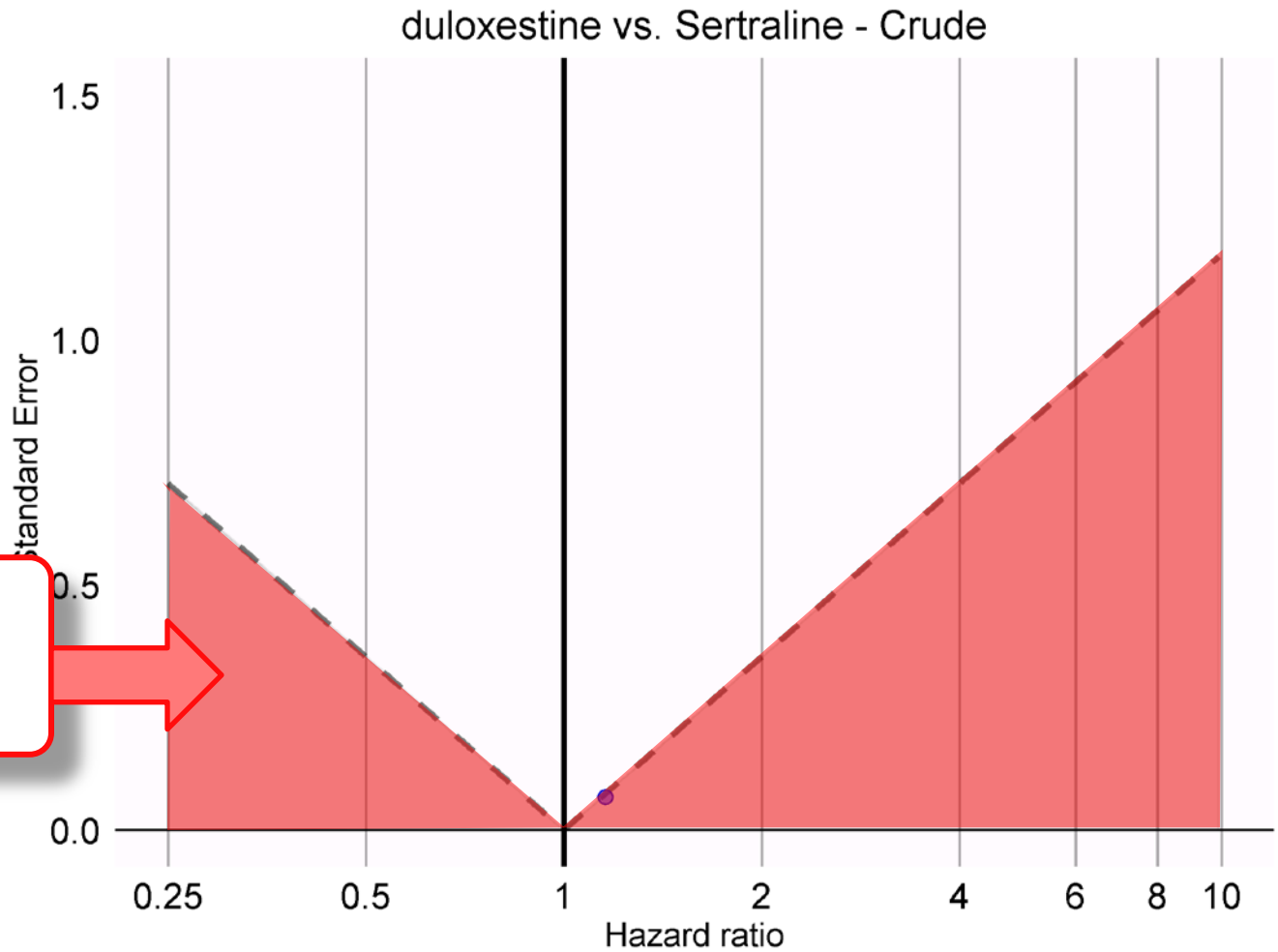


Negative control: ingrowing nail



Not statistically significant

Negative control: ingrowing nail

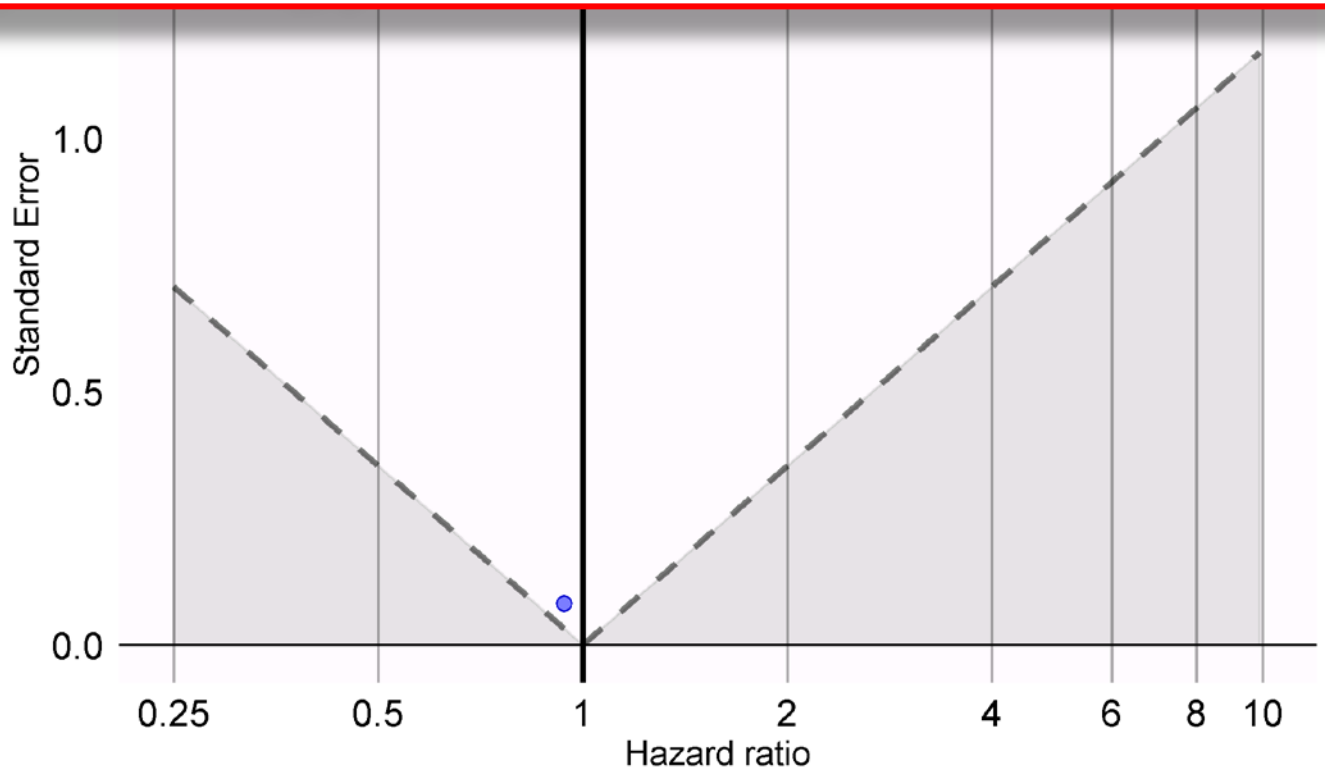


Statistically significant

Negative control: ingrowing nail

Adjusted estimate:

HR = 0.94 (0.80 – 1.10), p = 0.44



Depression – negative controls

Acariasis
Amyloidosis
Ankylosing spondylitis
Aseptic necrosis of bone
Astigmatism
Bell's palsy
Benign epithelial neoplasm of skin
Chalazion
Chondromalacia
Crohn's disease
Croup
Diabetic oculopathy
Endocarditis
Endometrial hyperplasia
Enthesopathy
Epicondylitis
Epstein-Barr virus disease

Ingrowing nail
Iridocyclitis
Irritable bowel syndrome
Lesion of cervix
Lyme disease
Malignant neoplasm of endocrine gland
Mononeuropathy
Onychomycosis
Osteochondropathy
Paraplegia
Polyp of intestine
Presbyopia
Pulmonary tuberculosis
Rectal mass
Sarcoidosis
Scar
Seborrheic keratosis

Generated with the help of LAERTES (see posters)

Hemangioma
Hodgkin's disease
Human papilloma virus infection
Hypoglycemic coma
Hypopituitarism
Impetigo

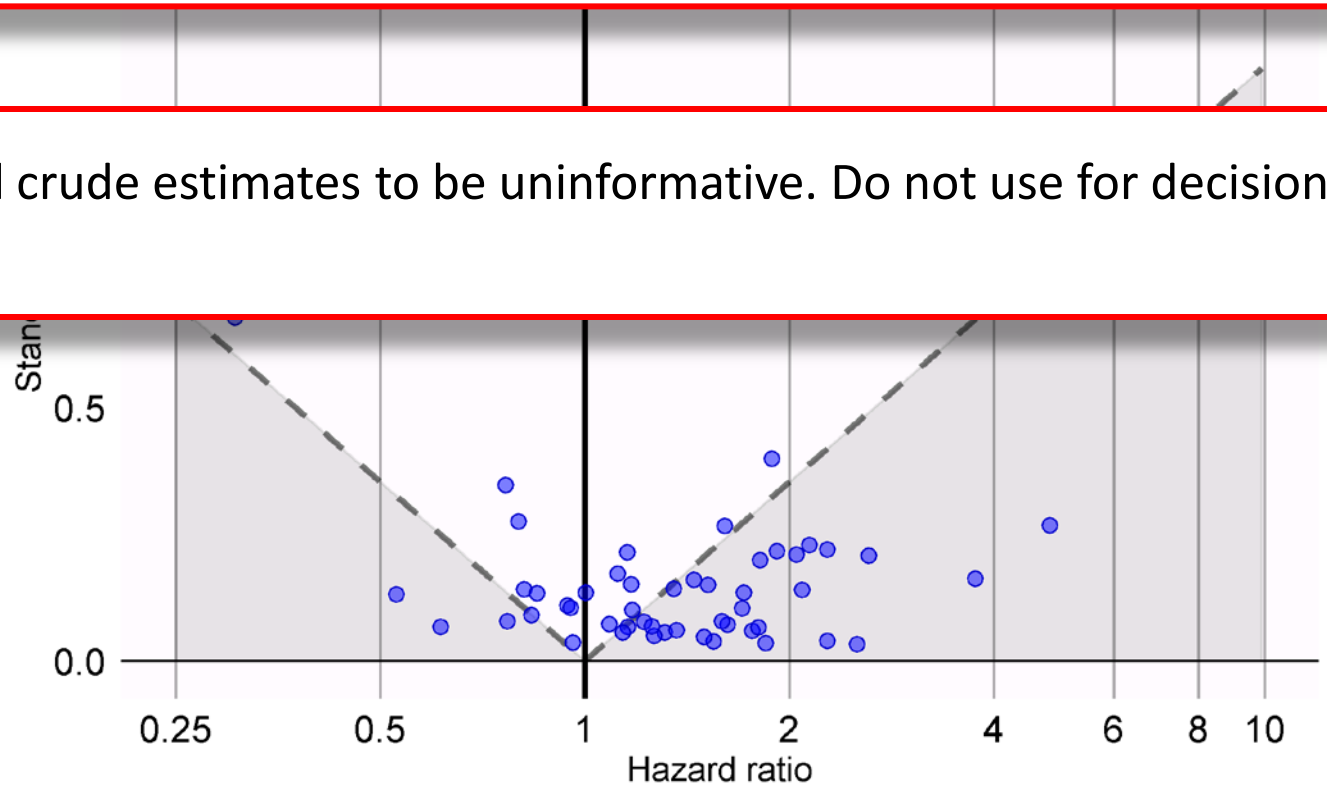
Tonsillitis
Toxic goiter
Ulcerative colitis
Viral conjunctivitis
Viral hepatitis
Visceroptosis

All negative controls - crude

We would expect 5% of negative controls to have $p < 0.05$

Instead, 68% has $p < 0.05$!

We found crude estimates to be uninformative. Do not use for decision making!

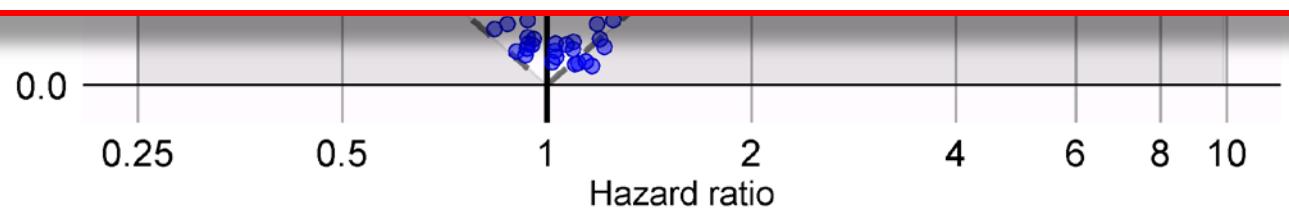


All negative controls - adjusted

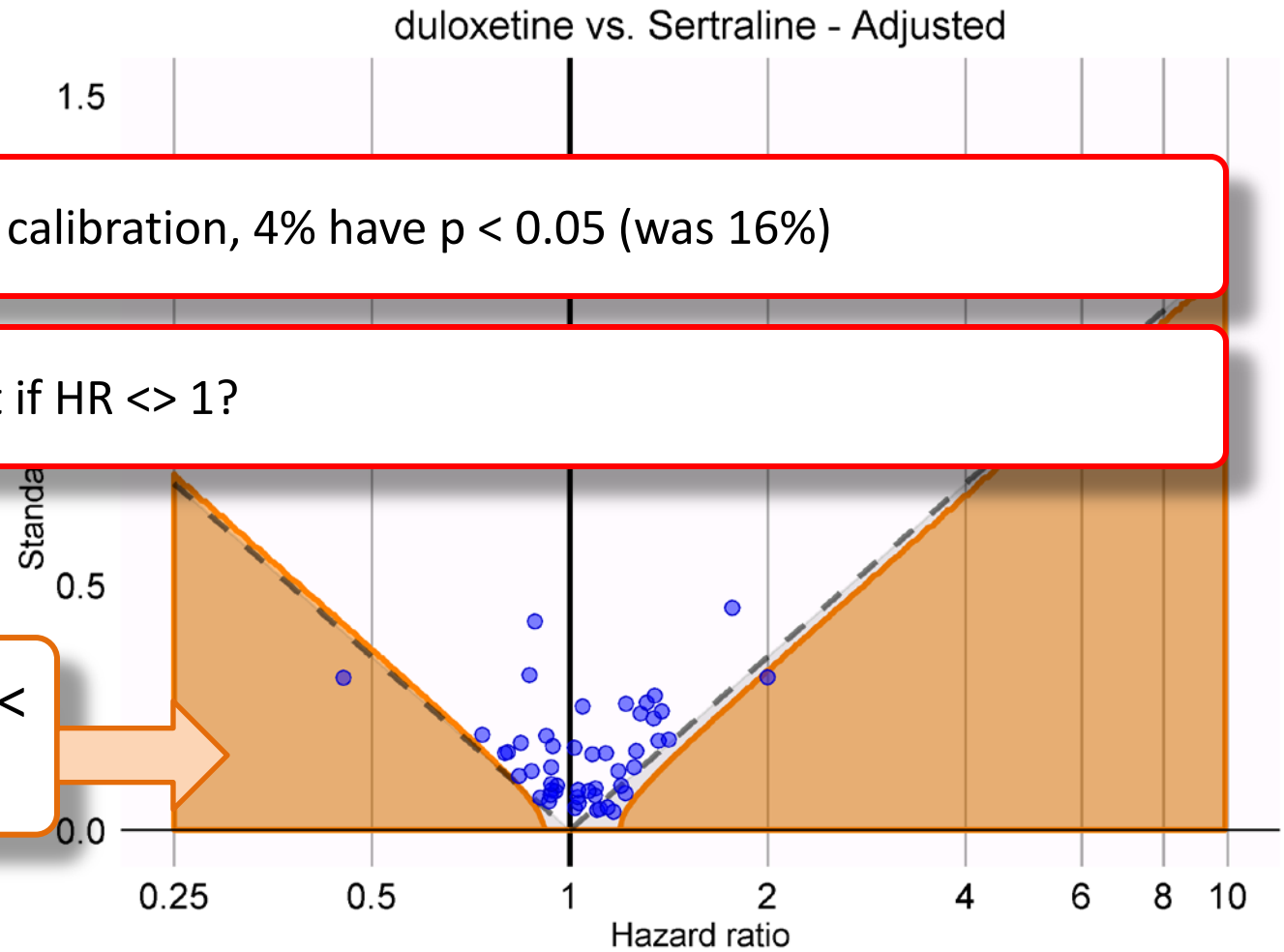
When using the propensity score, 16% have $p < 0.05$

In the past, we've shown you how you can perform p-value calibration:

- P-value represents probability of estimate when true RR = 1
- Negative controls provide empirical distribution of estimates when RR = 1
- Use empirical null distribution to compute calibrated p-value



P-value calibration



After calibration, 4% have p < 0.05 (was 16%)

What if HR <> 1?

Calibrated p < 0.05

Trouble with positive controls

- Often very few positive examples for a particular comparison
- Exact effect size never known with certainty (and depends on population)
- Doctors also know they're positive, and will change behavior accordingly

Drug Saf (2014) 37:655–659
DOI 10.1007/s40264-014-0198-z

CURRENT OPINION

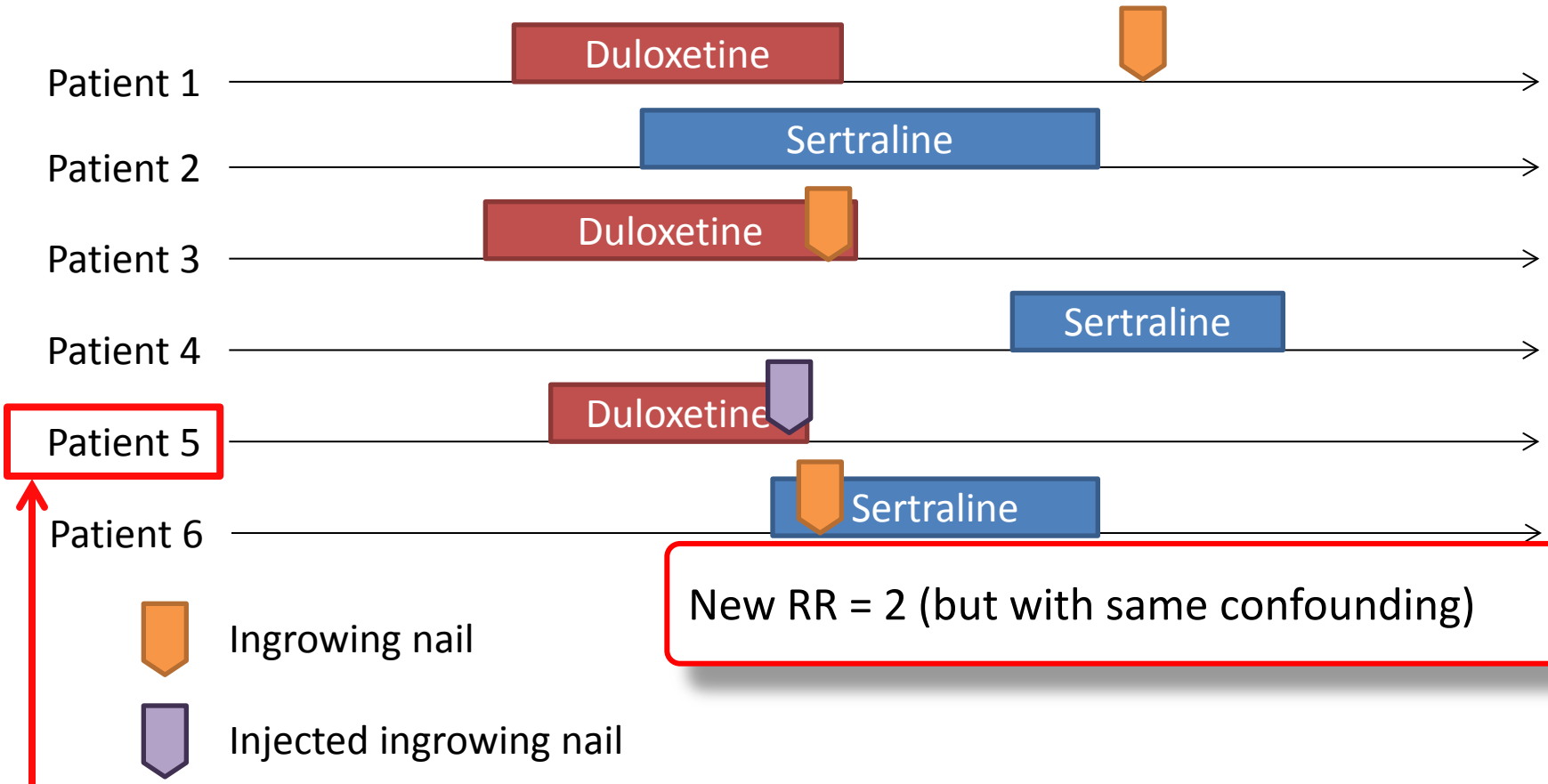
Zoo or Savannah? Choice of Training Ground for Evidence-Based Pharmacovigilance

G. Niklas Norén · Ola Caster · Kristina Juhlin · Marie Lindquist

Creating positive controls

- Start with negative controls: RR = 1
- Add simulated outcomes during exposure until desired RR is achieved
- Injected outcomes should behave like ‘real’ outcomes: preserve confounding structure by injecting outcomes for people at high risk

Creating positive controls

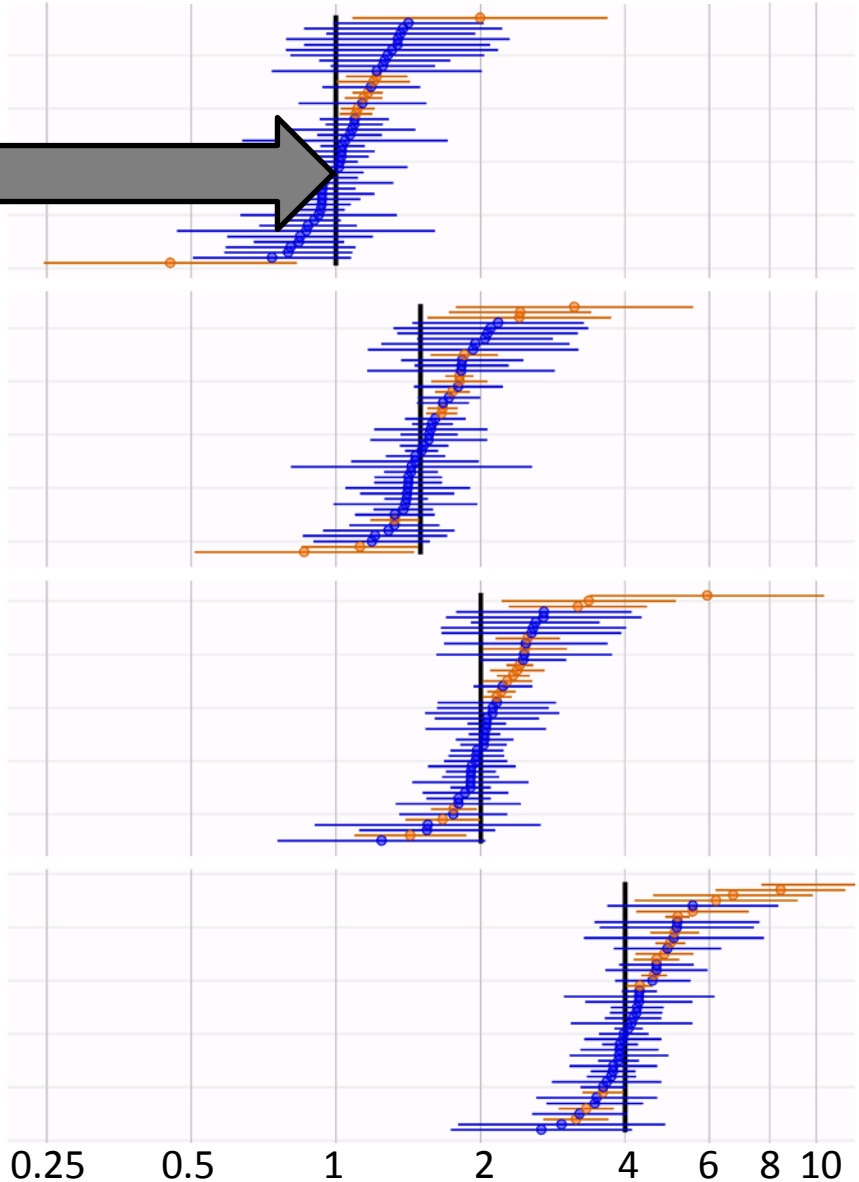


New RR = 2 (but with same confounding)

Predictive model of outcome indicates this is a high-risk patient

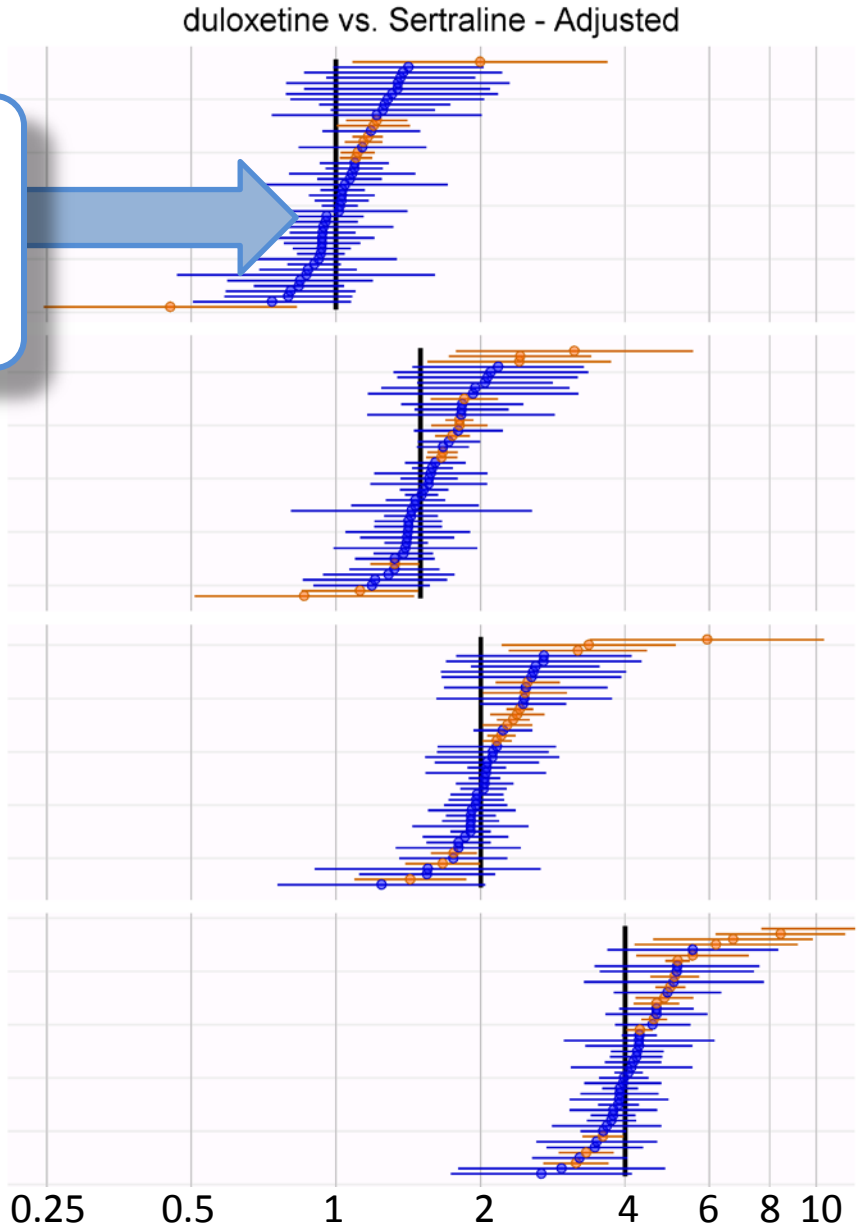
Estimated effects for positive controls

Black line indicates true hazard ratio



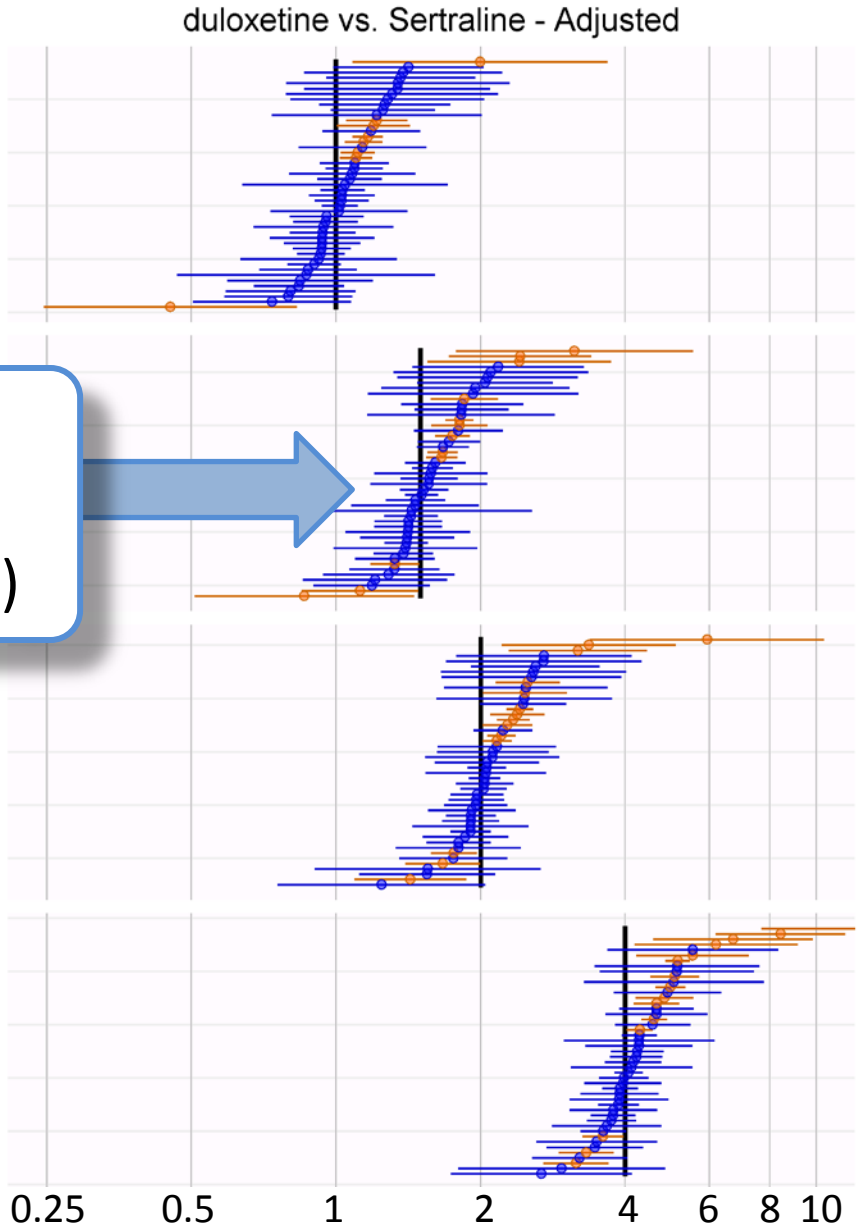
Estimating effects for positive controls

Ingrowing nail
True RR = 1
Estimated RR = 0.94 (0.80 – 1.10)

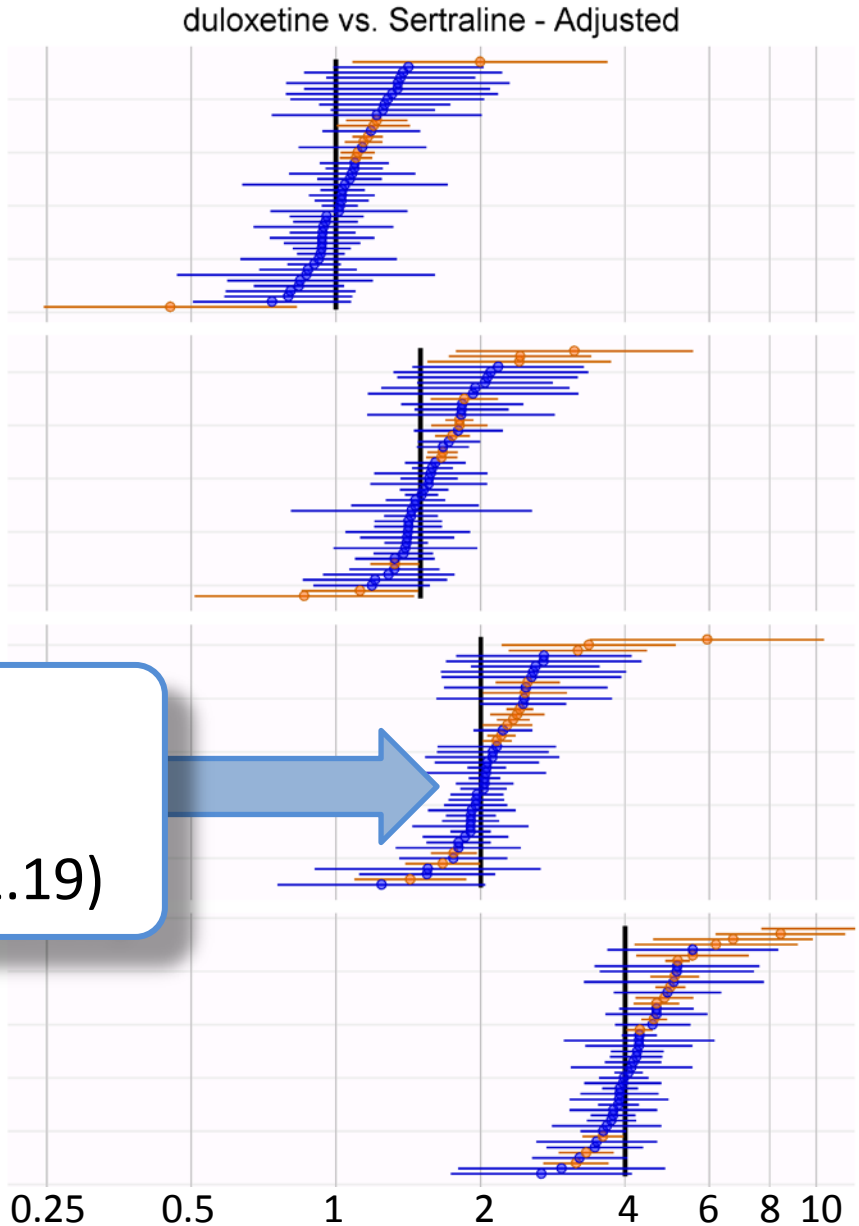


Estimating effects for positive controls

Ingrowing nail+
True RR = 1.5
Estimated RR = 1.47 (1.27 – 1.69)

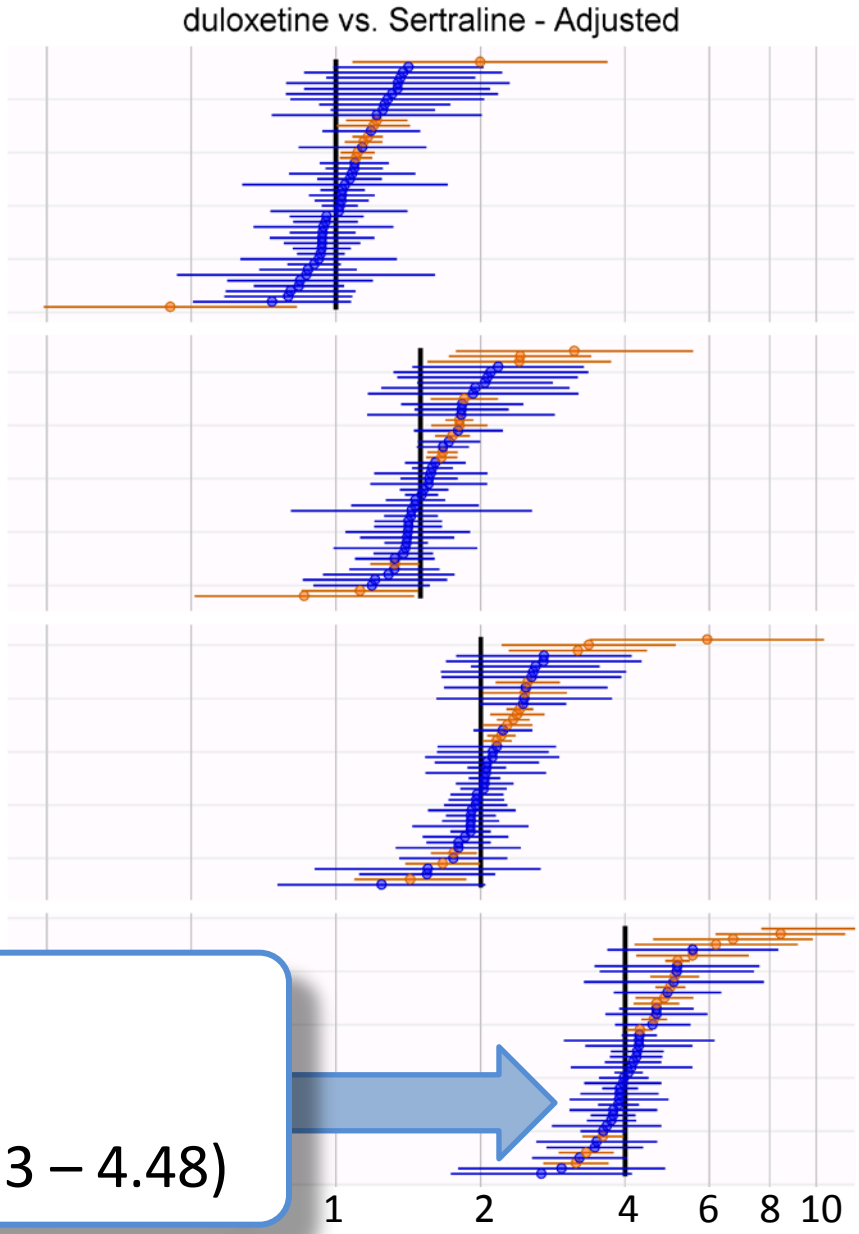


Estimating effects for positive controls



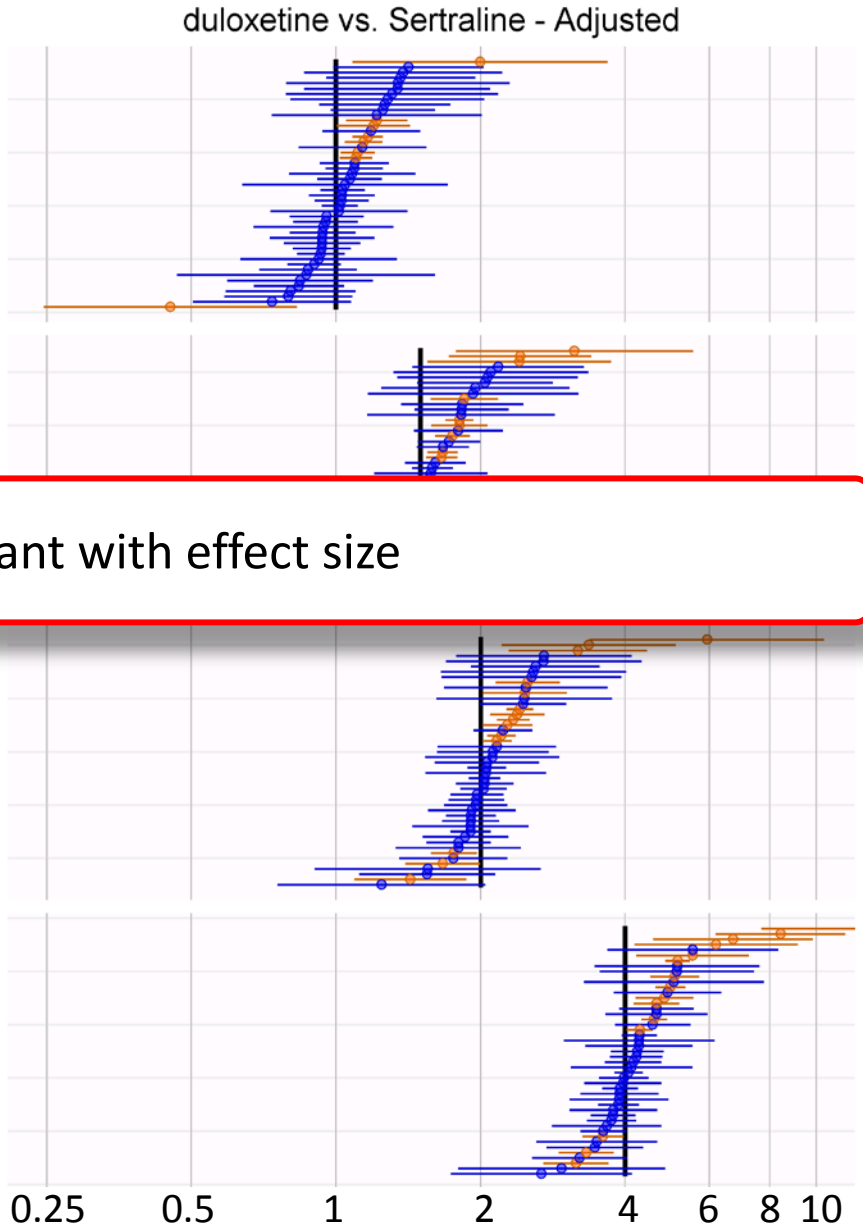
Ingrowing nail++
True RR = 2
Estimated RR = 1.91 (1.67 – 2.19)

Estimating effects for positive controls



Ingrowing nail+++
True RR = 4
Estimated RR = 3.89 (3.53 – 4.48)

Estimating effects for positive controls



Analysis suggests bias remains constant with effect size

Evaluating coverage of the CI

Coverage

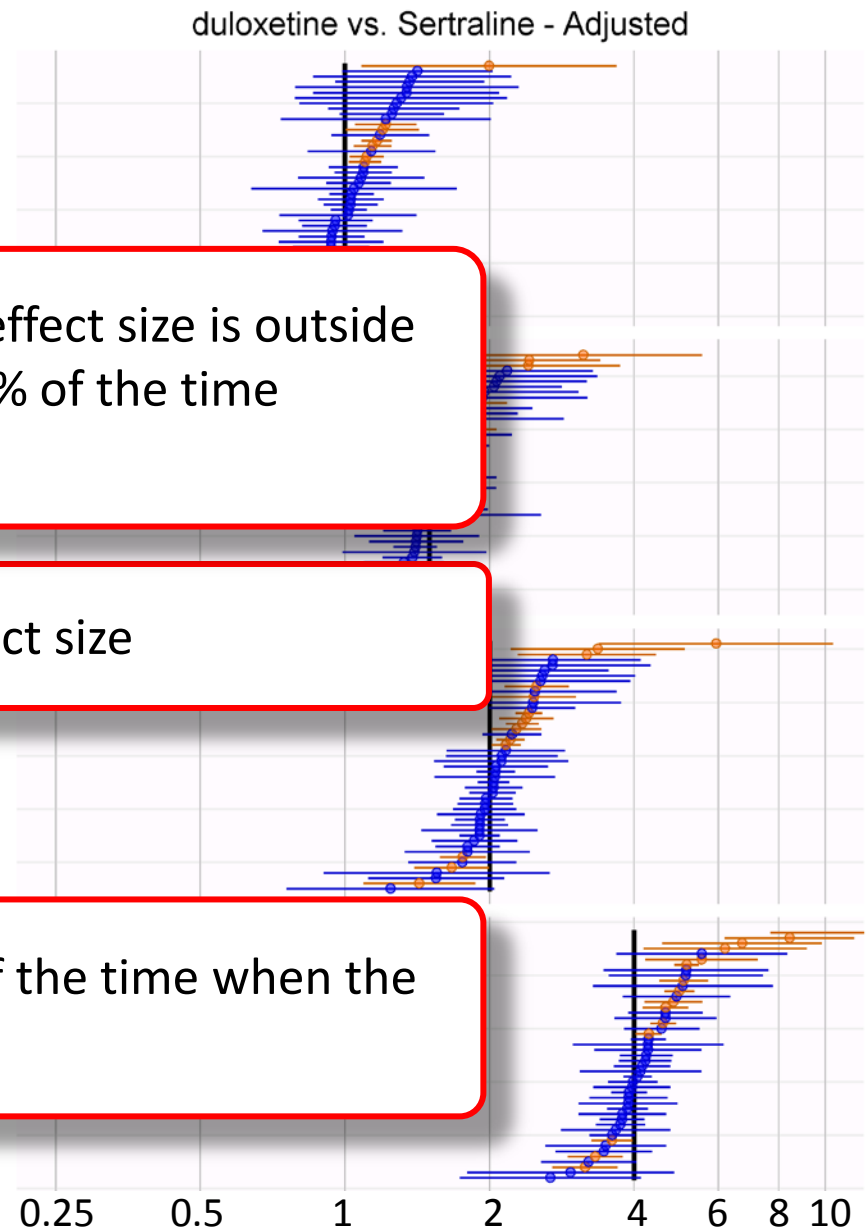
83%

Coverage of 83% means the true effect size is outside of the 95% confidence interval 17% of the time (when the true RR = 1)

Coverage decreases with true effect size

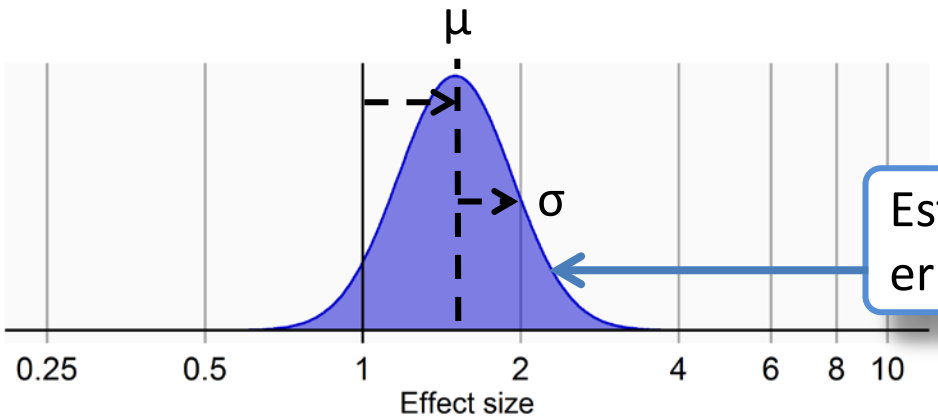
70%

Missing the true effect size 30% of the time when the true RR = 2!



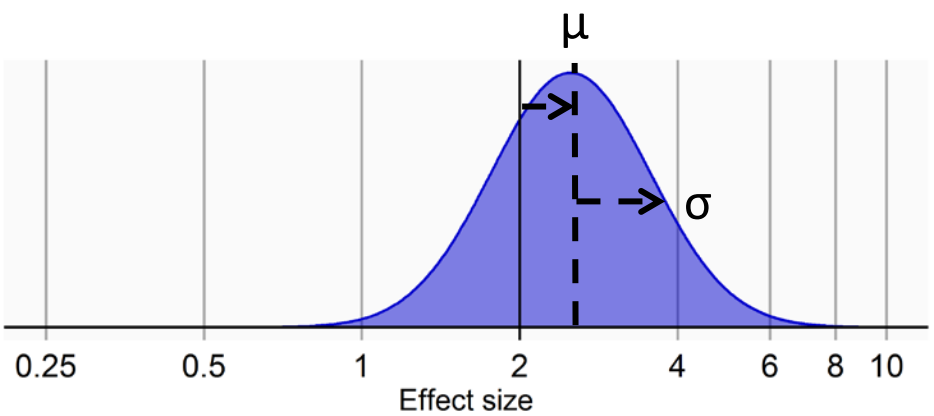
Confidence interval calibration

HR_{true}= 1



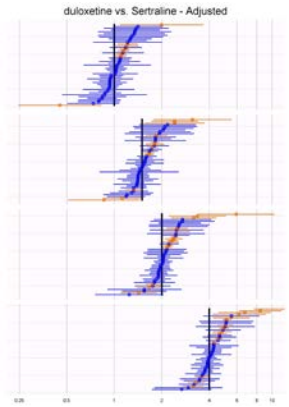
Estimated systematic error distribution

HR_{true}= 2



$$\mu = \alpha_\mu + \beta_\mu \log(HR_{true})$$
$$\sigma = \alpha_\sigma + \beta_\sigma \log(HR_{true})$$

Calibrating a confidence interval



$$\mu = 0.04 + 1.01 \log(\mathrm{HR}_{\mathrm{true}})$$
$$\sigma = 0.07 + 0.05 \log(\mathrm{HR}_{\mathrm{true}})$$

Uncalibrated

0.94 (0.80 - 1.10)

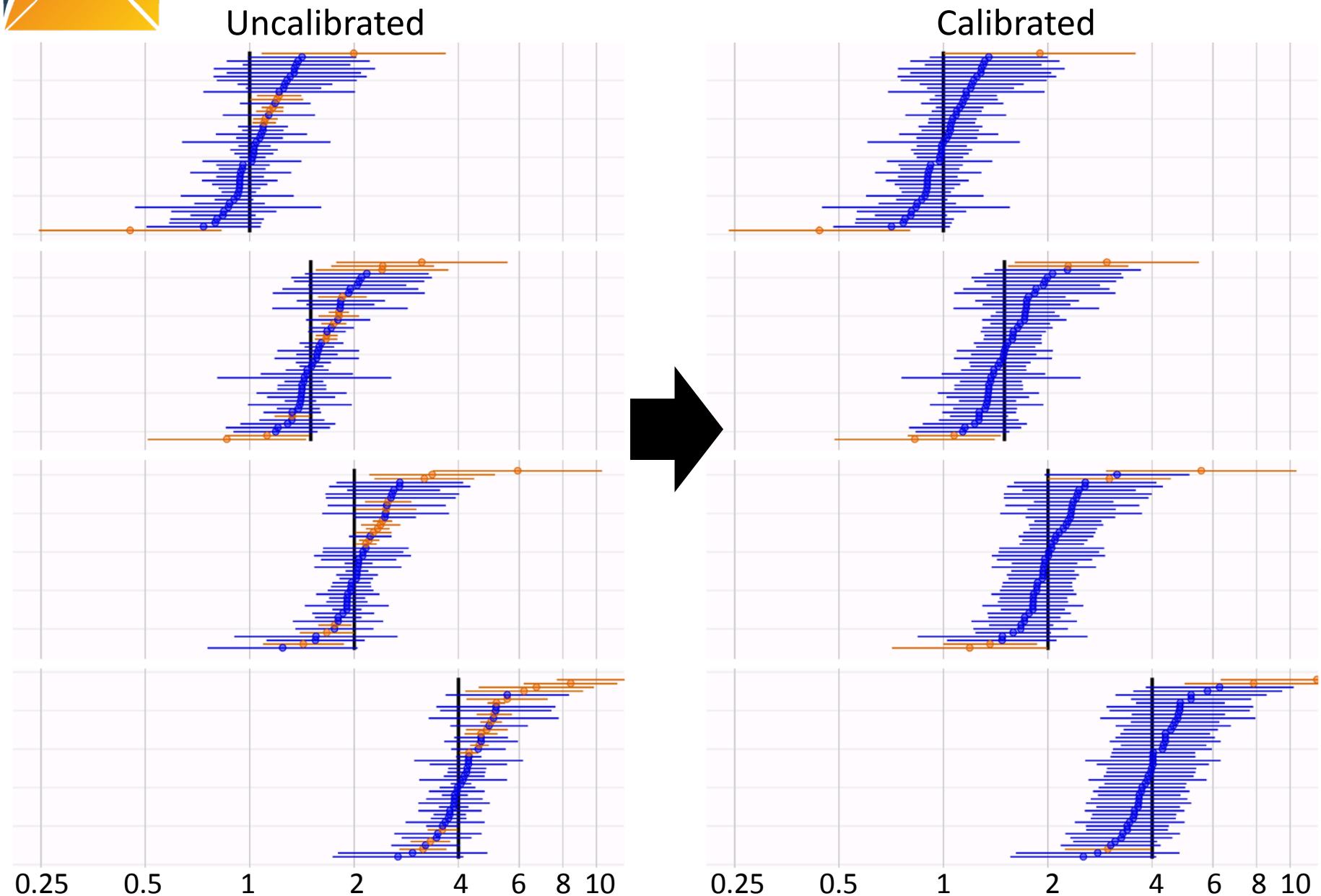
CI Calibration

Calibrated

0.90 (0.75 - 1.11)

Confidence intervals were too narrow, so made wider to get to nominal coverage

Confidence interval calibration



Confidence interval calibration

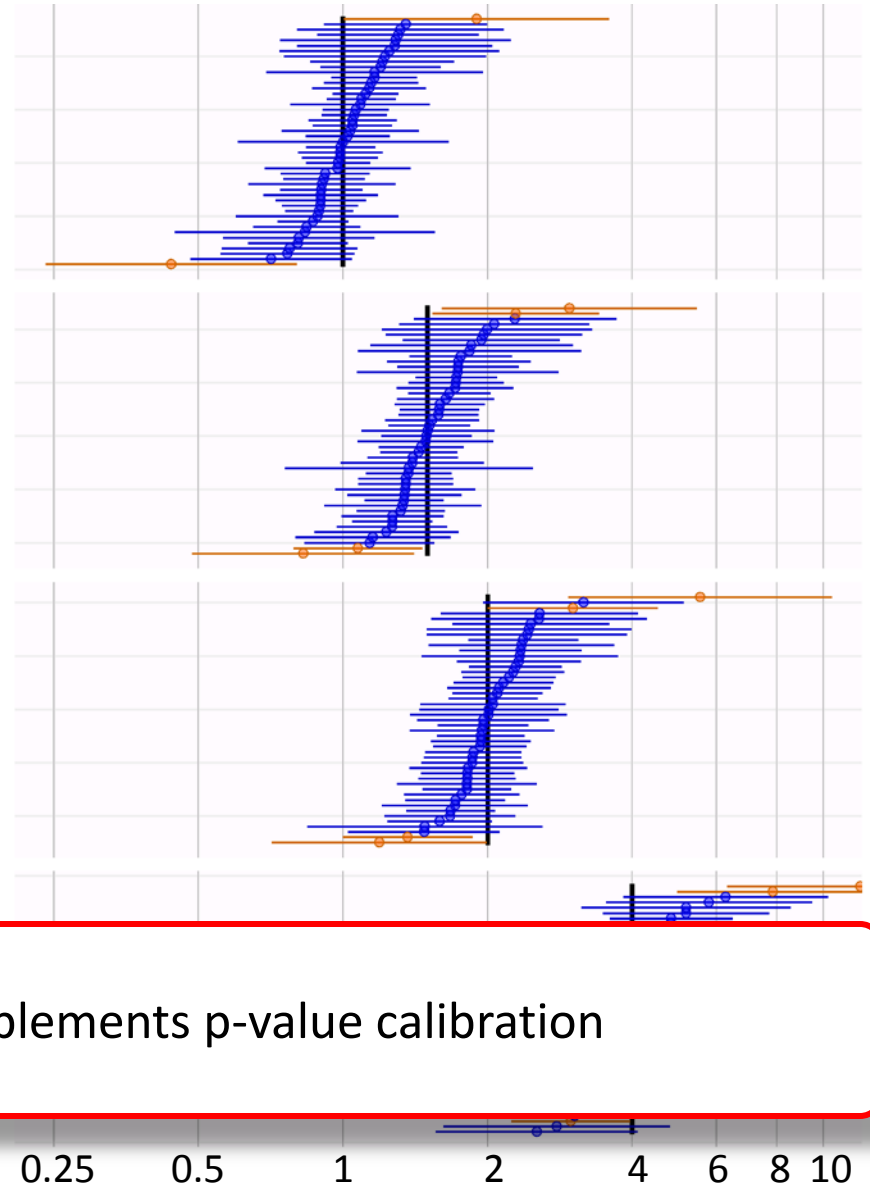
Coverage

Calibrated

96%

91%

91%



Confidence interval calibration complements p-value calibration

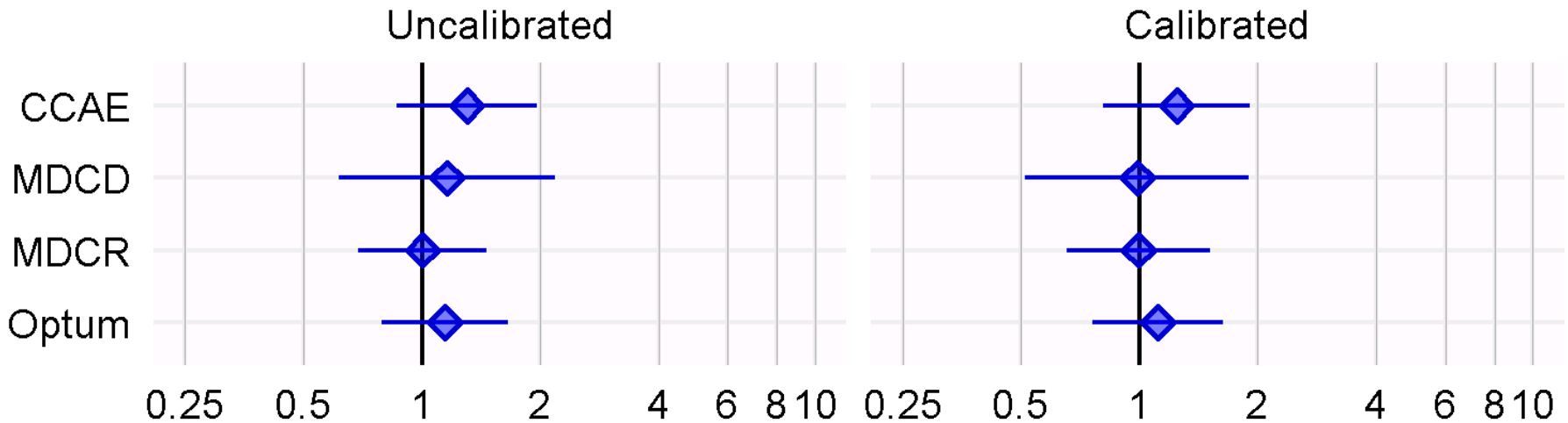
Current evidence for stroke

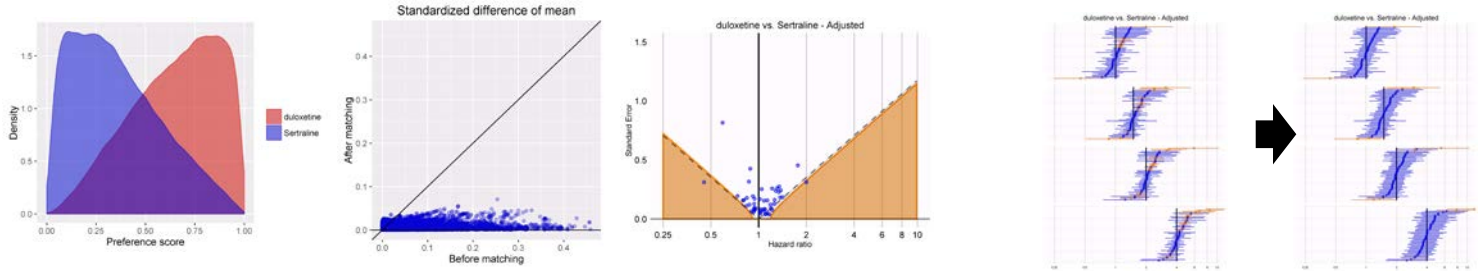
Result from Lee et al.

	Crude Hazard Ratio (95% CI)	*P*	Adjusted Hazard Ratio[a] (95% CI)	*P*
Main analyses				
SNRIs (n = 76,920) vs SSRIs (n = 582,650)				
Ischemic stroke	0.92 (0.83–1.02)	.12	1.01 (0.90–1.12)	.91

Proposed evidence for stroke

Duloxetine vs. Sertraline





Results are comparable to Lee et al., but we provide the context to interpret the results

OHDSI recommendations for evidence evaluation

- ✓ Produce standard diagnostics
  - E.g. for cohort studies diagnose the propensity score distribution, covariate balance, etc.
- ✓ Include negative controls
  - Estimate the error when the null is true
- ✓ Create positive controls
  - Estimate the error when RR > 1
- ✓ Calibrate p-value and confidence intervals
  - Restoring nominal characteristics

Population-level effect estimation

- How do we share evidence to inform decision making?

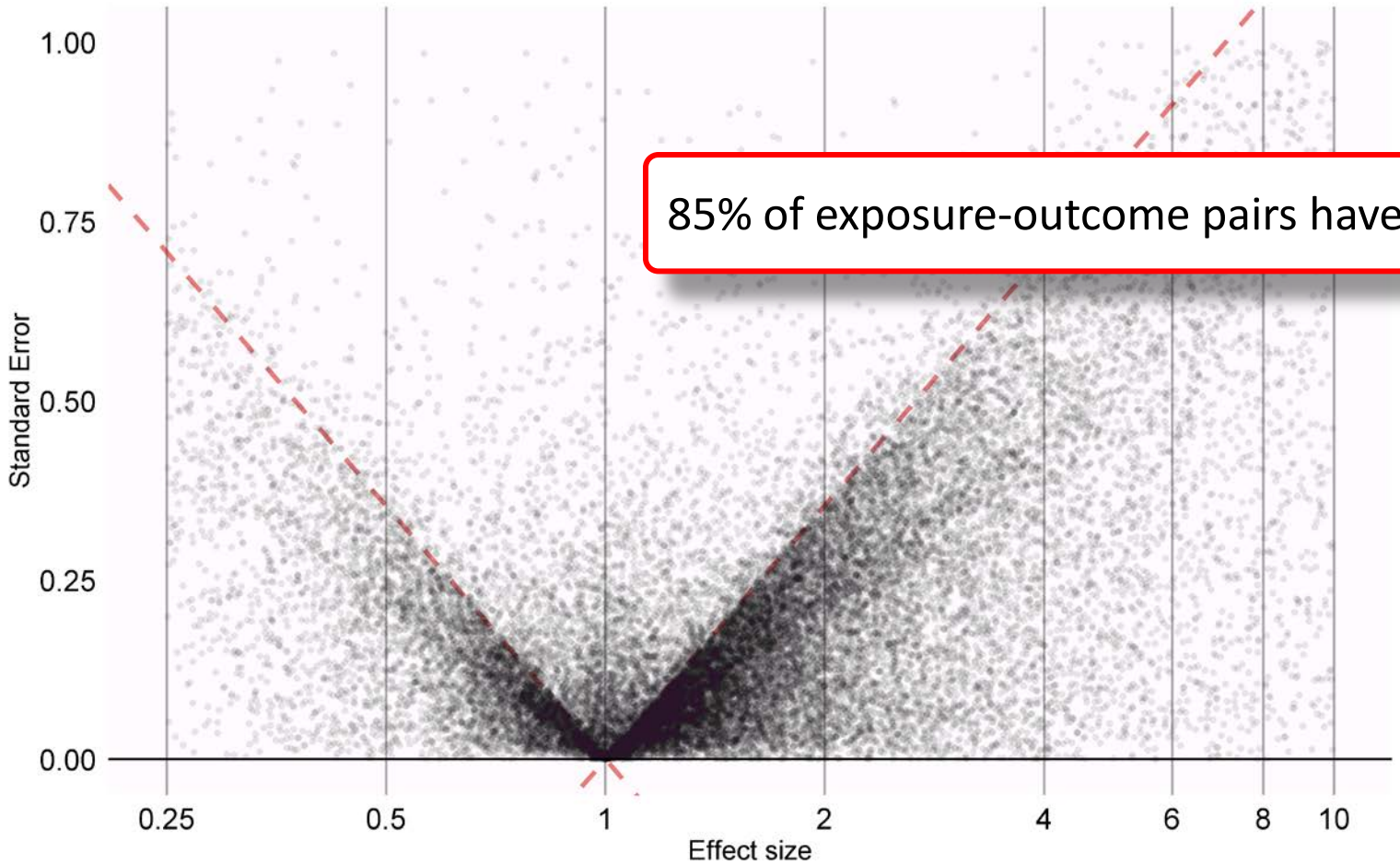
Evidence dissemination

- Traditionally, this evidence is disseminated through the scientific literature

- How well does that work?

Automated extraction of effect sizes from literature

Observational research results in literature



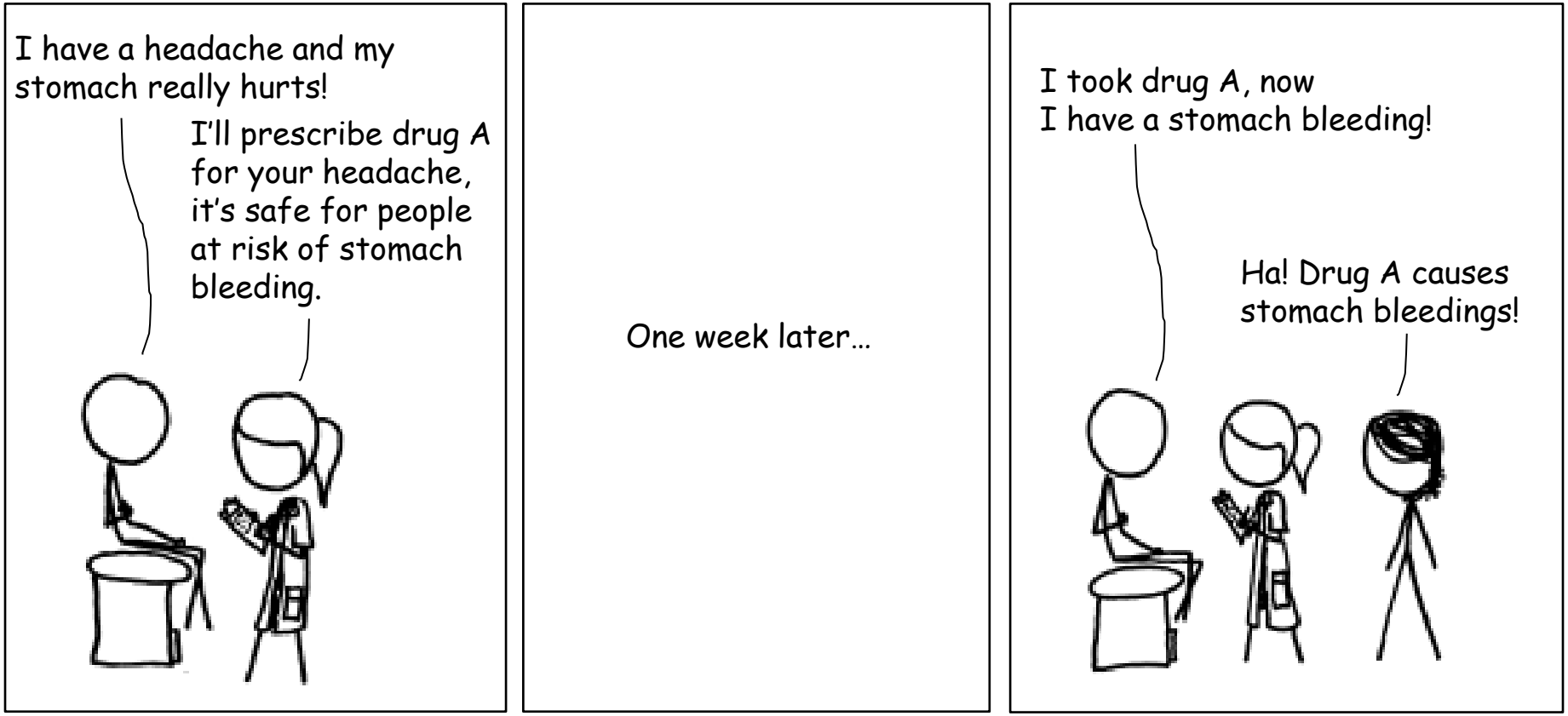
85% of exposure-outcome pairs have $p < 0.05$

29,982 estimates
11,758 papers

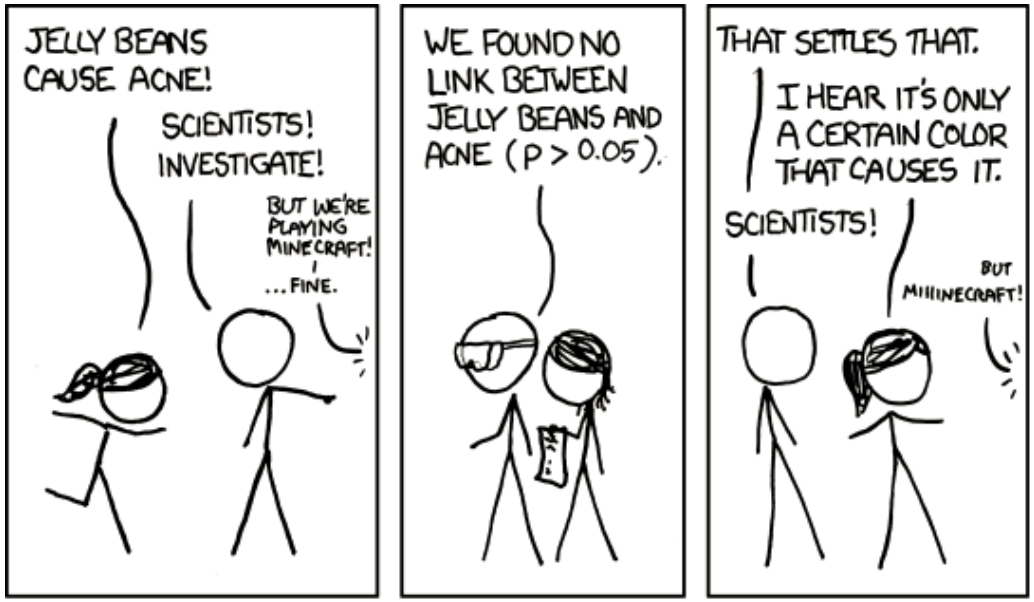
What went wrong?

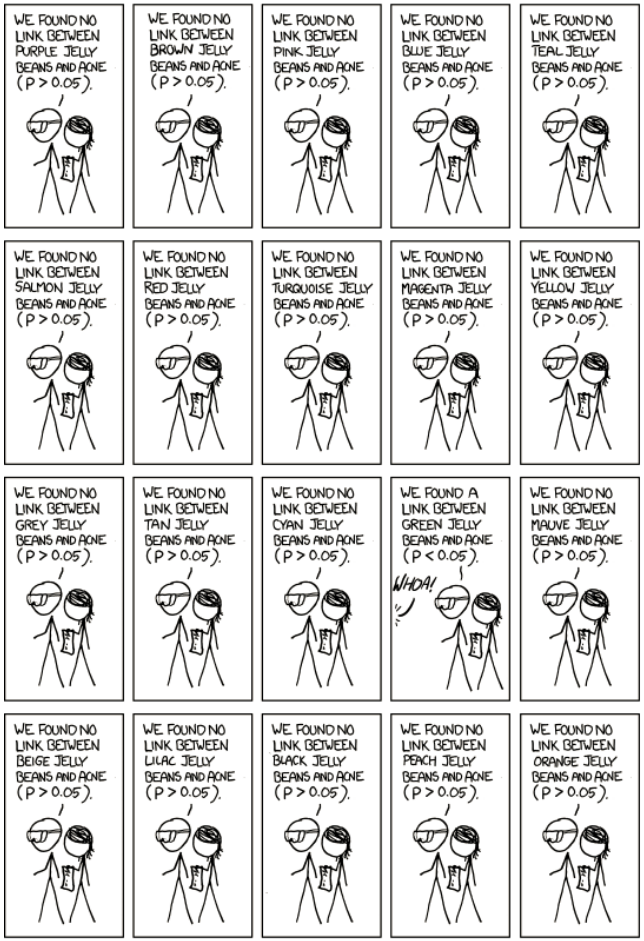
- Observational study bias
- Publication bias
- P-hacking

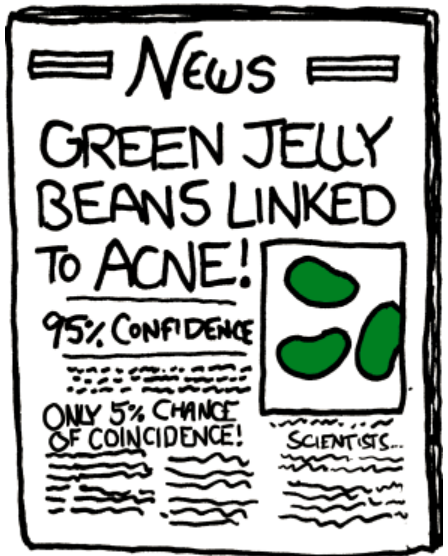
Observational study bias



Publication bias

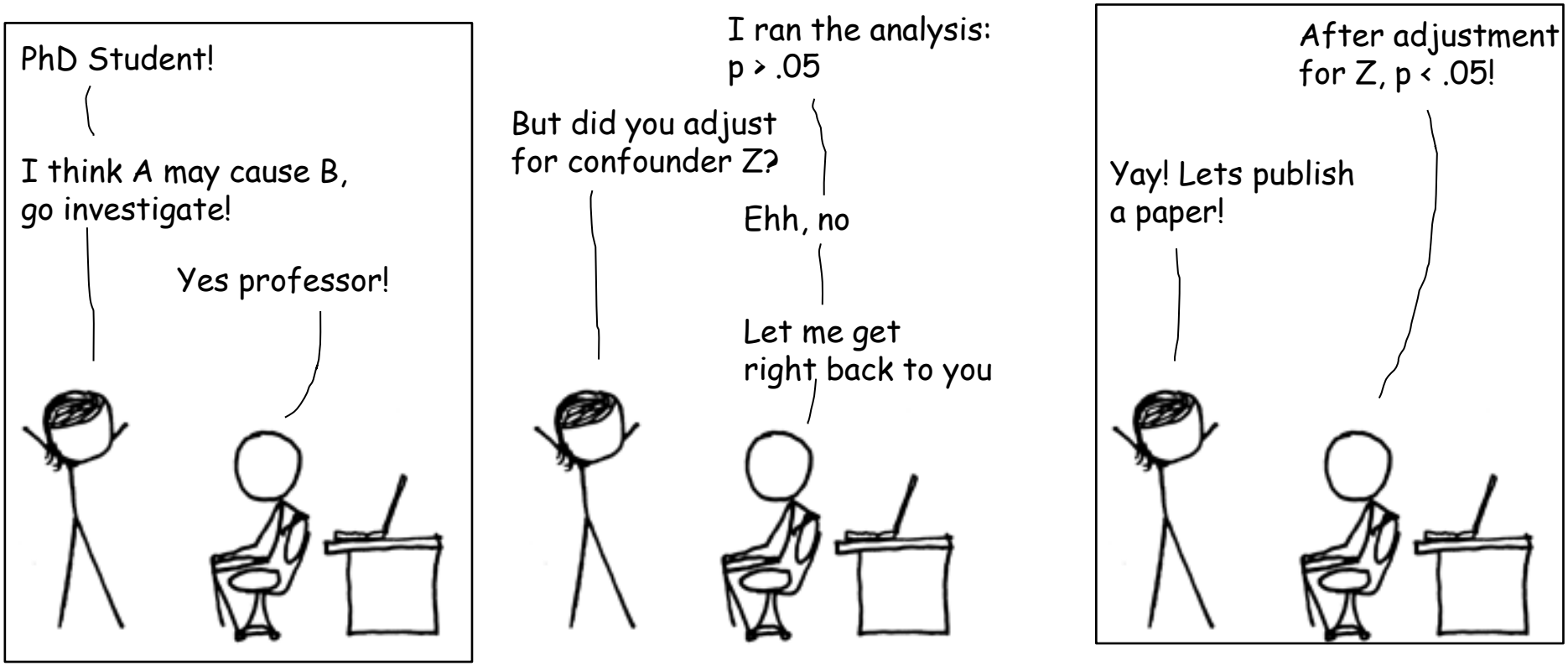






http://xkcd.com/882/

P-hacking


PhD Student!
I think A may cause B, go investigate!
Yes professor!
But did you adjust for confounder Z?
I ran the analysis: p > .05
Ehh, no
Let me get right back to you
Yay! Lets publish a paper!
After adjustment for Z, p < .05!

A solution?

Stop doing one study at a time!

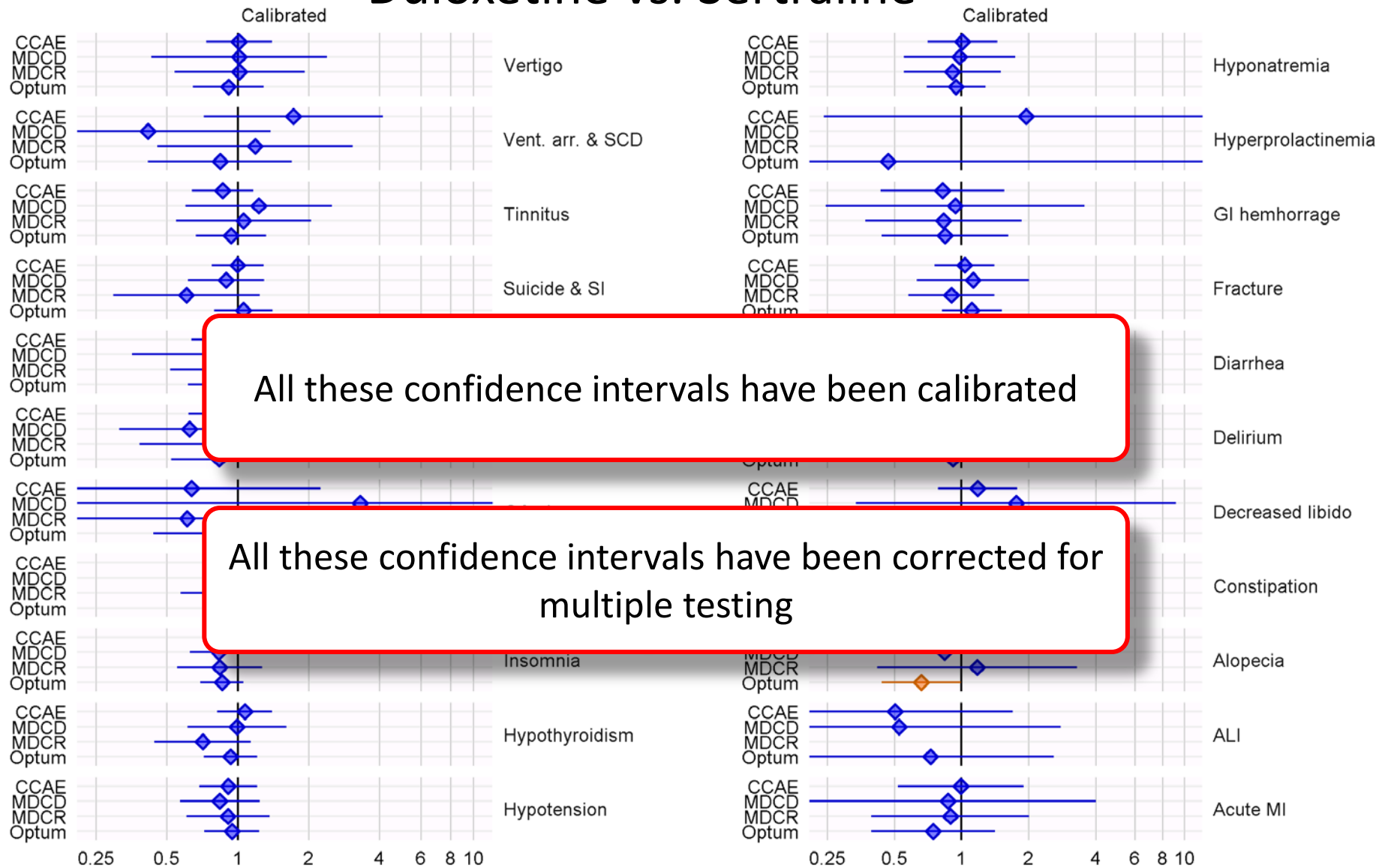
What if we considered all outcomes?

Duloxetine vs. Sertraline for these 22 outcomes:

Acute liver injury	Hypotension
Acute myocardial infarction	Hypothyroidism
Alopecia	Insomnia
Constipation	Nausea
Decreased libido	Open-angle glaucoma
Delirium	Seizure
Diarrhea	Stroke
Fracture	Suicide and suicidal ideation
Gastrointestinal hemorrhage	Tinnitus
Hyperprolactinemia	Ventricular arrhythmia and sudden cardiac death
Hyponatremia	Vertigo

All outcomes

Duloxetine vs. Sertraline

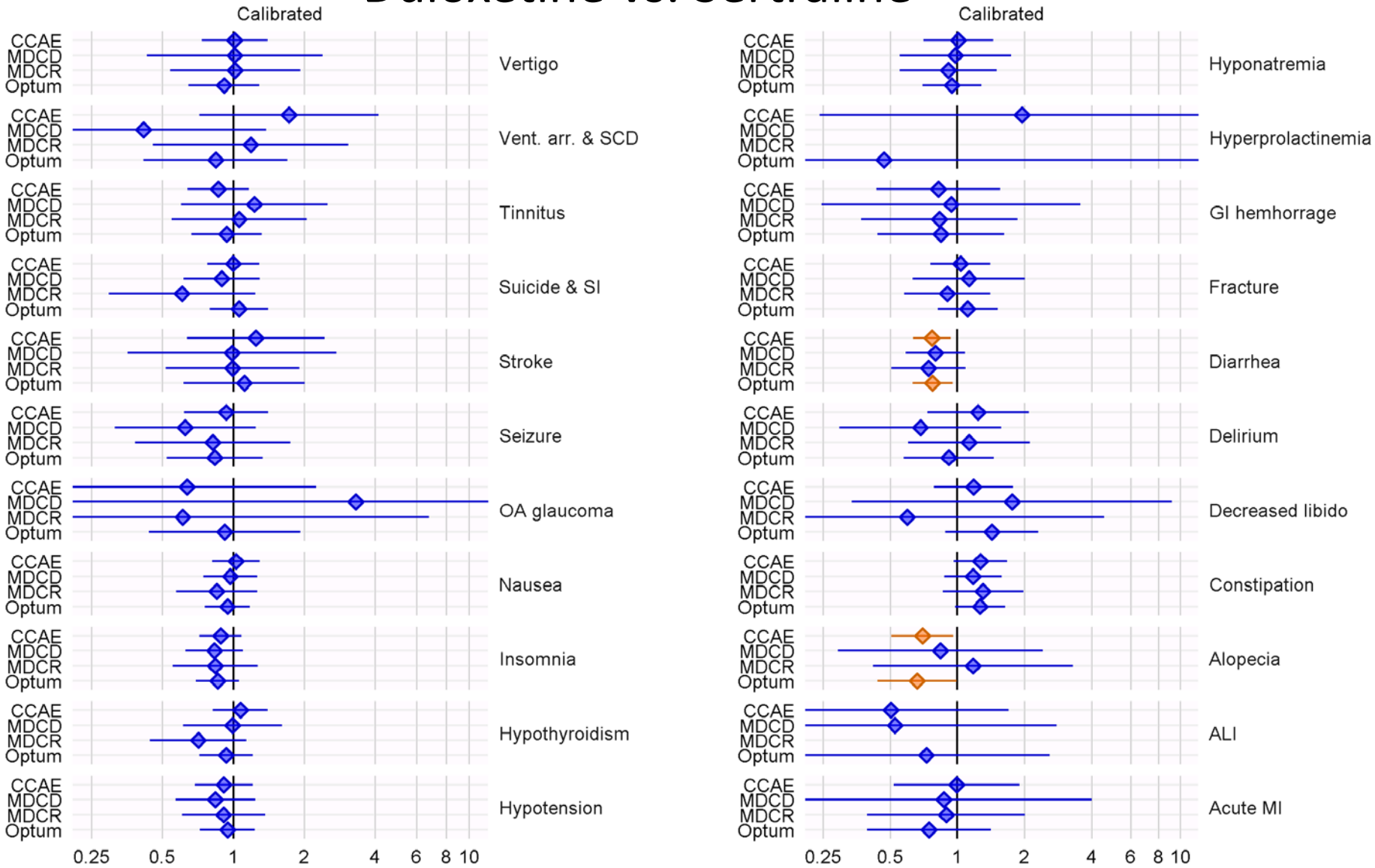


All these confidence intervals have been calibrated

All these confidence intervals have been corrected for multiple testing

All outcomes

Duloxetine vs. Sertraline



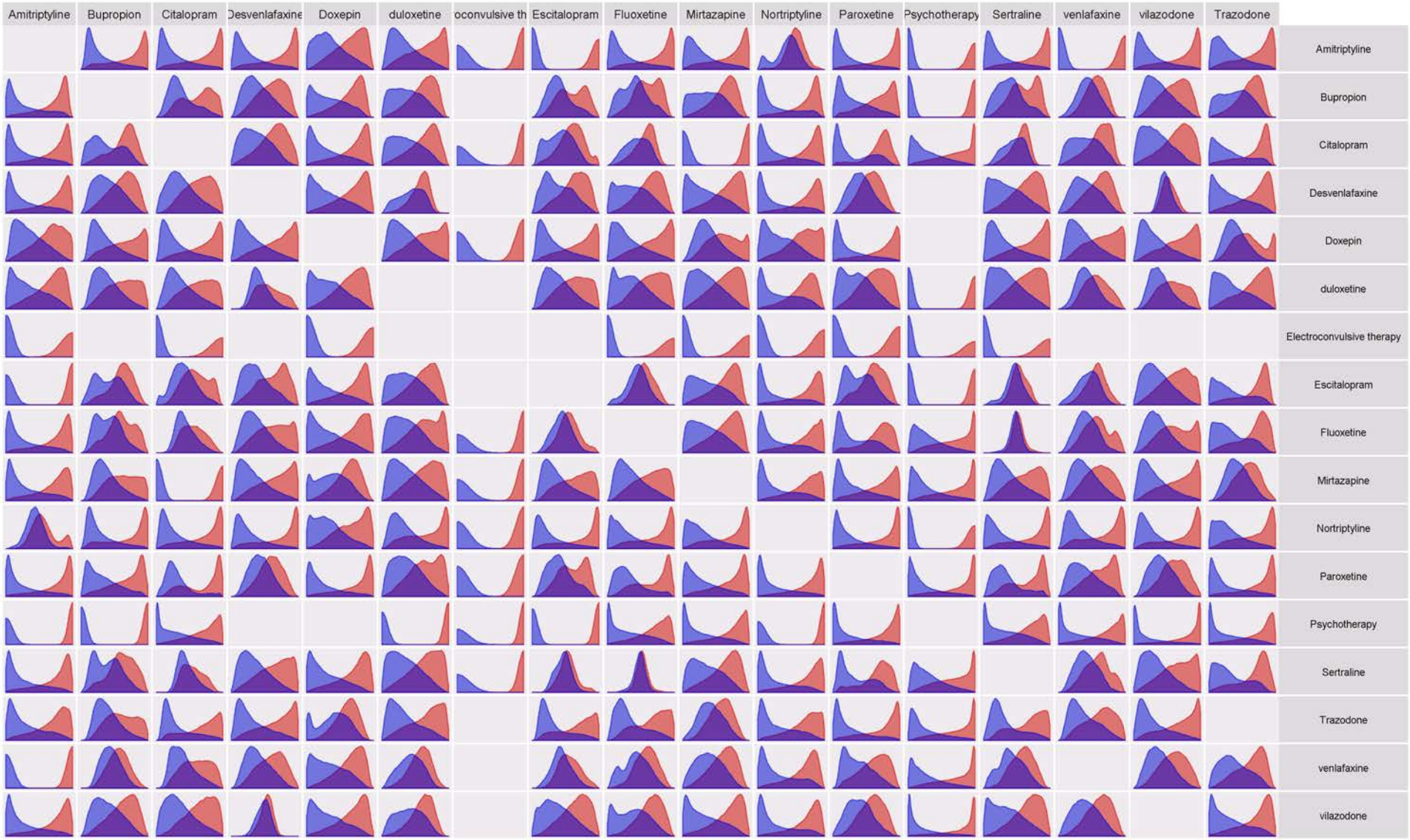
What if we consider all treatments?

Type	Class	Treatment
Drug	Atypical	Bupropion
Drug	Atypical	Mirtazapine
Procedure	ECT	Electroconvulsive therapy
Procedure	Psychotherapy	Psychotherapy
Drug	SARI	Trazodone
Drug	SNRI	Desvenlafaxine
Drug	SNRI	duloxetine
Drug	SNRI	venlafaxine
Drug	SSRI	Citalopram
Drug	SSRI	Escitalopram
Drug	SSRI	Fluoxetine
Drug	SSRI	Paroxetine
Drug	SSRI	Sertraline
Drug	SSRI	vilazodone
Drug	TCA	Amitriptyline
Drug	TCA	Doxepin
Drug	TCA	Nortriptyline

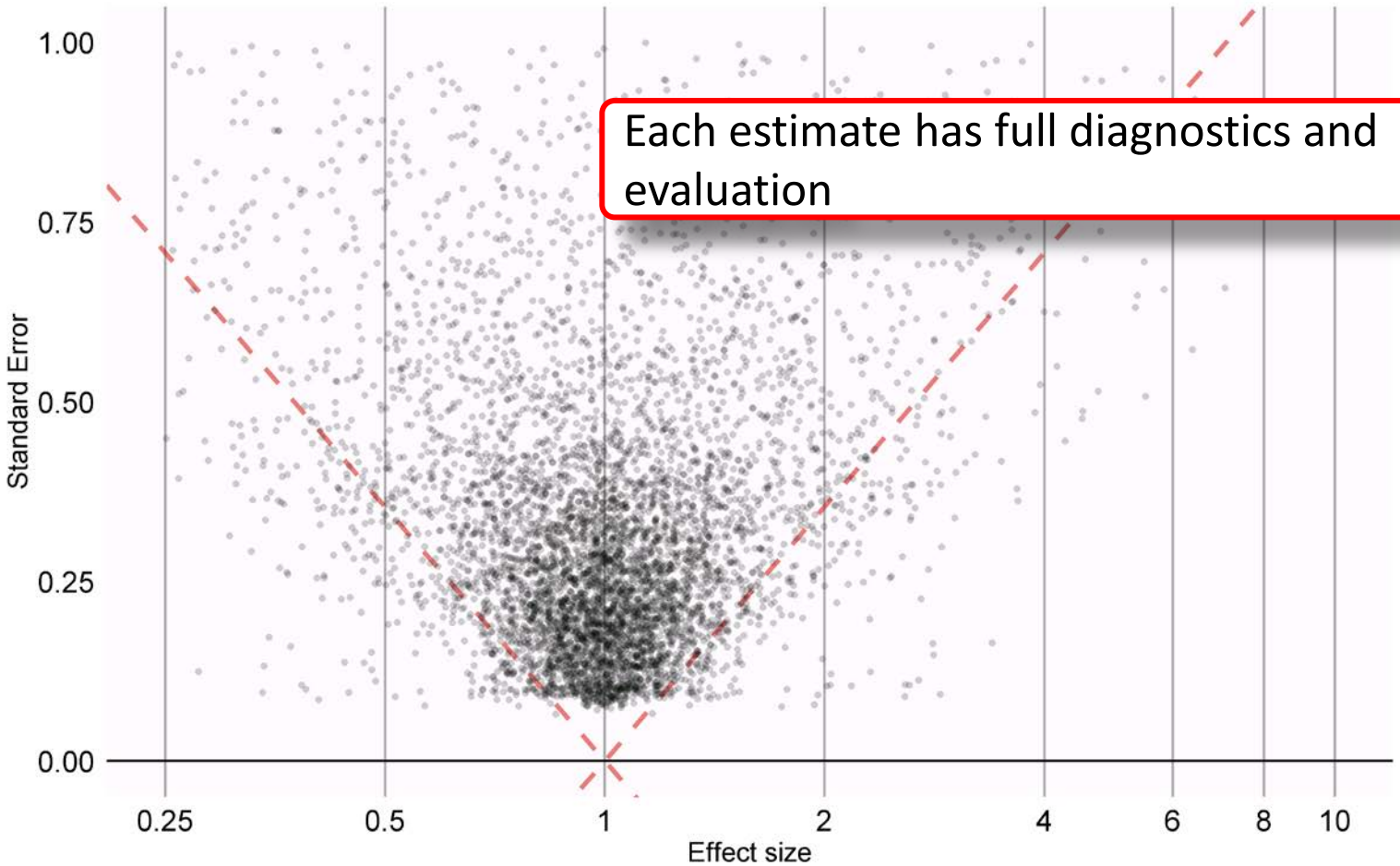
Large-scale estimation for depression

- 17 **treatments**
- 17 * 16 = 272 comparisons
- 22 **outcomes**
- 272 * 22 = 5,984 effect size estimates
- 4 **databases** (Truven CCAE, Truven MDCD, Truven MDCR, Optum)
- 4 * 5,984 = **23,936 estimates**

Propensity models for all comparisons (Truven CCAE)

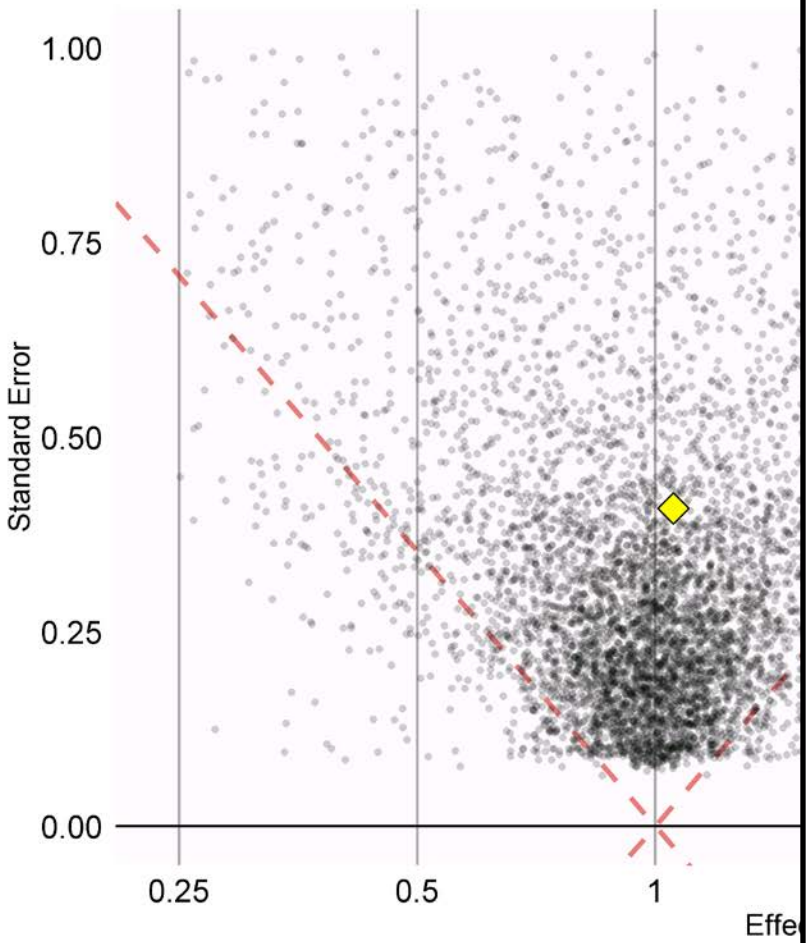


Large-scale estimation for depression



Each estimate has full diagnostics and evaluation

Example 1

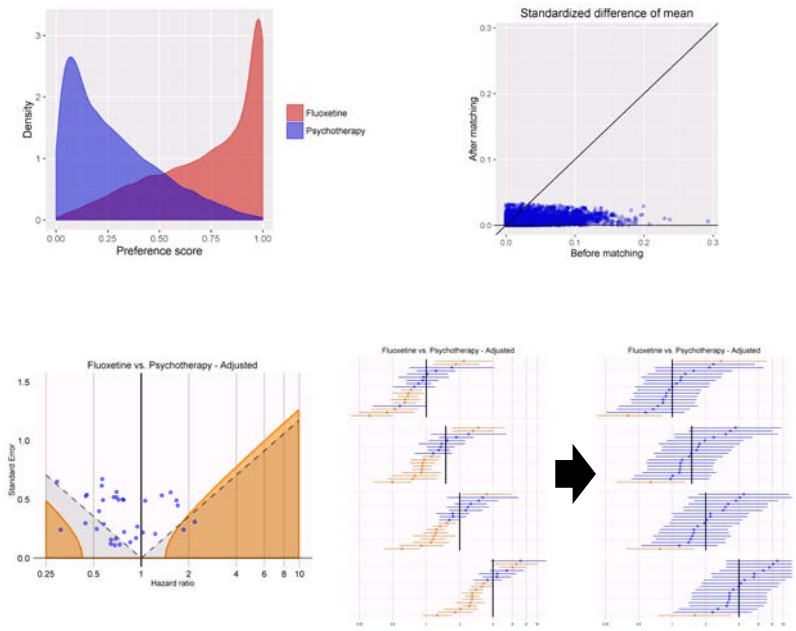


Fluoxetine vs. psychotherapy

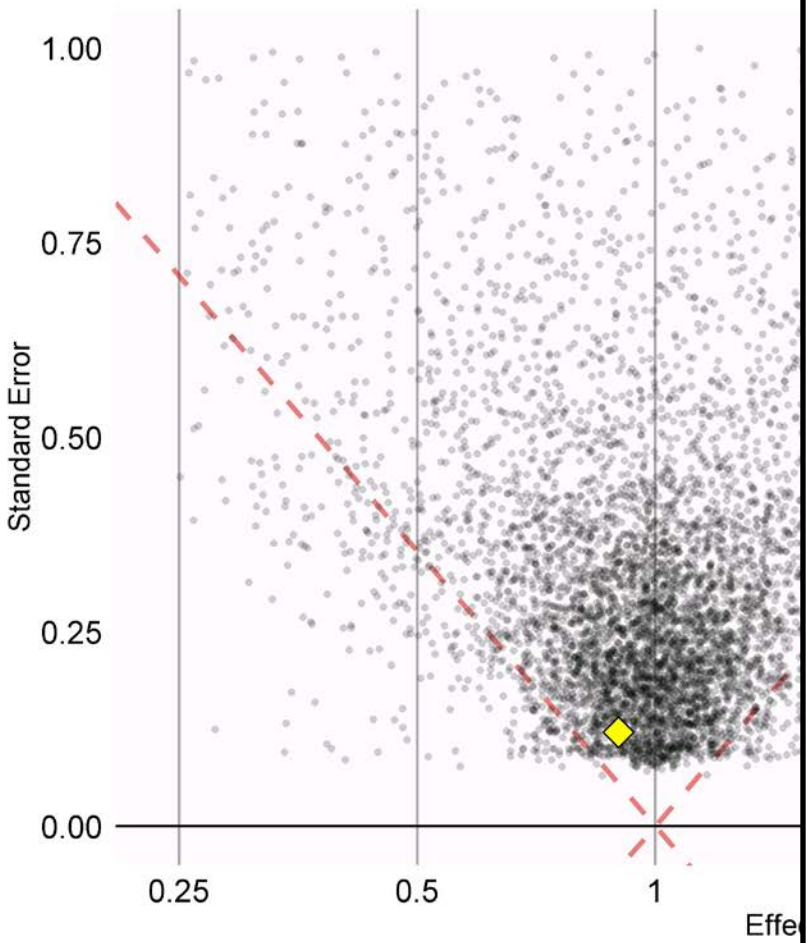
Suicide ideation

Database: Truven MDCR

Calibrated HR = 1.05 (0.51 – 2.51)



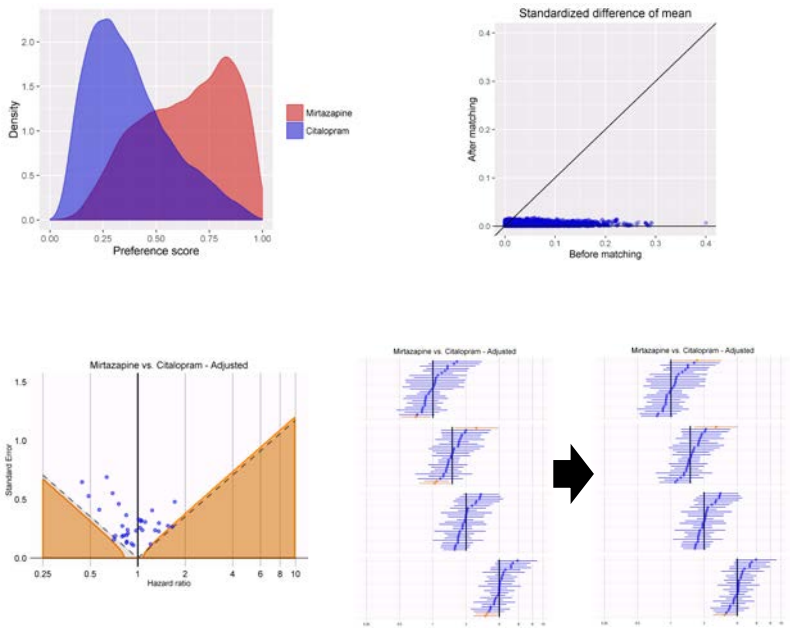
Example 2



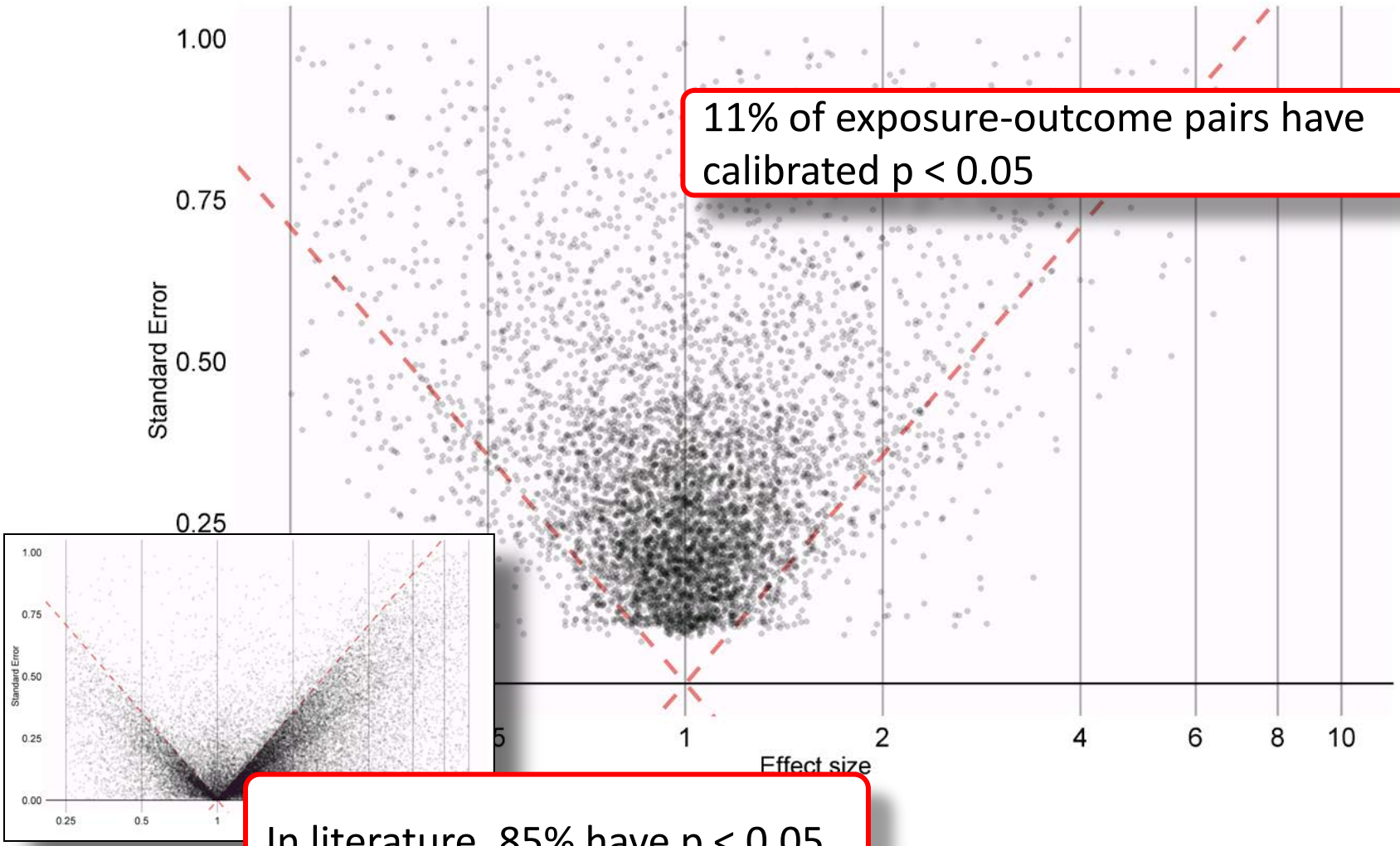
Mirtazapine vs. Citalopram
Constipation

Database: Truven MDCD

Calibrated HR = 0.90 (0.70 – 1.12)



Estimates are in line with expectations



11% of exposure-outcome pairs have calibrated p < 0.05

In literature, 85% have p < 0.05

Large-scale estimation for depression

- Each estimate produced with same rigor, and could be published as a paper
  - Propensity score adjustment
  - Cox regression
  - Calibrated using negative and positive controls
  - …
- Not “data-mining”!
  - Results should be interpreted considering multiple testing
  - This can’t be done for literature

OHDSI recommendations for evidence dissemination

- ✓ Address observation study bias

  Addressed by adjusting for confounding, and **verifying** bias was addressed. Disseminate your diagnostics and evaluations.

- ✓ Address publication bias

  Avoided by showing all tests that were performed, not just those with $p < 0.05$

- ✓ Address p-hacking

  Very hard to fine-tune analysis to one specific result

Population-level effect estimation

- Write and share protocol
- Open source study code
- Use validated software
- Replicate across databases

- Produce standard diagnostics
- Include negative controls
- Create positive controls
- Calibrate confidence interval and p-value

- Don't provide only the effect estimate
- Also share protocol, study code, diagnostics and evaluation
- Produce evidence at scale

Building the LHC of observational research?


Join the journey