OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Evaluating Use of Methods for Adverse Event Under Surveillance (EUMAEUS)

Why EUMAEUS?

1) The rapid rollout of COVID-19 vaccines makes it increasingly critical to perform large-scale evaluations of vaccine safety using real-world evidence.

2) Estimate the comparative performance (bias, precision, timeliness) of the case-control, cohort, historical rate, and self-controlled methods for vaccine safety.

Literature Review

Lana Lai

on behalf of the EUMAEUS task force

Types of Study Designs

Study Design	Description	Advantages	Disadvantages	Clinical Applications
Case-control	• Comparison of cases vs. non-cases from the same source population from the same time-period	• Uses small data sample from entire cohort, cost efficient • Use matching to control for time-varying confounders	• Confounding by indication • Selection bias • Misclassification of exposure	• Autism spectrum disorders & various vaccines • Inflammatory bowel disease (IBD) & MMR vaccine • Guillain-Barré syndrome (GBS) & H1N1 vaccine

Types of Study Designs

Study Design	Description	Advantages	Disadvantages	Clinical Applications
Cohort	- Comparison of incidence rate ratio of adverse events between vaccinated vs. unvaccinated population	- Easy to implement – large amount of data available - Use matching / stratification to control for potential confounders	- Confounding by indication - Misclassification of exposure	- Intussusception & rotavirus vaccine - Autism spectrum disorders & various vaccines
Historical Rate (Comparator) Cohort	- Comparison between observed incidence of adverse events vs. expected incidence based on historical data	- Greater statistical power to detect rare adverse events - Improved timeliness in detecting potential safety signals	- Temporal confounders (e.g. seasonality, changing trends in detection of adverse events & variation in diagnostic criteria over time)	- Pediatric vaccines - Tdap vaccine - HPV vaccine - H1N1 vaccine

Types of Study Designs

Study Design	Description	Advantages	Disadvantages	Clinical Applications
Self-Controlled Case Series (SCCS) / Self-Controlled Risk Interval (SCRI)	• Comparison between incidence rates in exposed time periods vs. incidence rates of self-matched unexposed time periods • SCCS: Cases only • SCRI: Vaccinated population only	• Adjust for time-invariant confounders • SCCS: Multiple occurrences of independent events within an individual can be assessed • SCRI: Less susceptible to misclassification of exposure	• Time-varying confounding (e.g. age, seasonality) • Reverse causality bias	• Guillain-Barré syndrome (GBS) & H1N1 vaccine • Autism spectrum disorders & various vaccines • Seizures & various vaccines

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

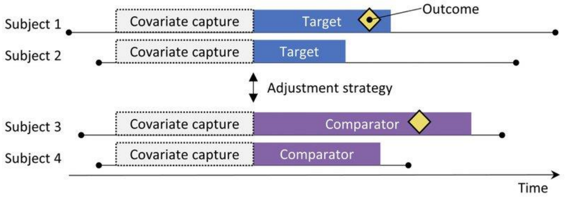
Overview of the EUMAEUS Experiment Design

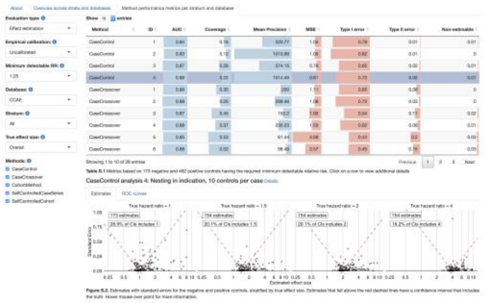
Marc Suchard

on behalf of the EUMAEUS task force

EUMAUES is an empirical benchmark study

Builds on our prior work evaluation of comparative (drug) effectiveness and safety methods published in *Harvard Data Science Review*




About
Overview across strata and databases
Method performance metrics per stratum and database
Evaluation type: Effect estimation
Empirical calibration: Uncalibrated
Minimum detectable RR: 1.25
Database: CCAE
Stratum: All
True effect size: Overall
Methods:
CaseControl
CaseCrossover
CohortMethod
SelfControlledCaseSeries
SelfControlledCohort
Show 10 entries

Method	ID	AUC	Coverage	Mean Precision	MSE	Type I error	Type II error	Non-estimable
CaseControl	1	0.84	0.16	320.77	1.04	0.78	0.01	0.01
CaseControl	2	0.83	0.12	1010.89	1.06	0.82	0.01	0
CaseControl	3	0.87	0.26	574.15	0.78	0.65	0.02	0.01
CaseControl	4	0.89	0.22	1014.49	0.81	0.72	0.02	0.01
CaseCrossover	1	0.86	0.35	250	1.11	0.65	0.06	0
CaseCrossover	2	0.89	0.25	206.46	1.06	0.73	0.03	0
CaseCrossover	3	0.87	0.46	153.2	1.65	0.64	0.11	0.02
CaseCrossover	4	0.86	0.37	235.23	1.55	0.62	0.06	0.01
CaseCrossover	5	0.85	0.53	61.44	4.08	0.44	0.2	0.03
CaseCrossover	6	0.88	0.52	98.49	2.57	0.45	0.16	0.03

Showing 1 to 10 of 28 entries Previous 1 2 3 Next

Table 5.1 Metrics based on 173 negative and 482 positive controls having the required minimum detectable relative risk. Click on a row to view additional details.

CaseControl analysis 4: Nesting in indication, 10 controls per case Details

Figure 5.2 Estimates with standard errors for the negative and positive controls, stratified by true effect size. Estimates that fall above the red dashed lines have a confidence interval that includes the truth. Hover mouse over point for more information.

Harvard Data Science Review • 2.1

How Confident Are We About Observational Findings in Health Care: A Benchmark Study

Martijn J. Schuemie, M. Soledad Cepede, Marc A. Suchard, Jianxiao Yang, Yuxi Tian Alejandro Schuler, Patrick B. Ryan, David Madigan[1], George Hripcsak

[1]Professor of Statistics, Columbia University

Published on: Jan 31, 2020

DOI: 10.1162/99608f92.147cc28e

To systematically evaluate the

performance of methods

to reliably

identify vaccine safety signals

in

real-world settings

Vaccine safety surveillance methods

Reduce systematically to **four** components:

- Construction of a ***counterfactual*** (“expected count” without vaccination)
- A ***time-at-risk*** when safety events can occur
- The ***test-statistic*** to estimate, and
- A ***decision rule*** to classify signals from non-signals

Counterfactual construction

- Case-control
  - How often are patients with events vaccinated?
- Contemporary non-user comparator cohort method
  - How often do events occur to similar unvaccinated patients?
  - Some variants: ***anchoring*** (or *not*) on healthcare visit; ***matching*** (or *not*) on age + sex
- Historical rates
  - How often did events occur to other patients in the past?
  - Some variants: ***anchoring***; ***stratifying*** (or *not*) on age + sex
- Self-control case series
  - How often did/do events occur in the same patients at different times?

Note: 17 total variations drawn from the literature

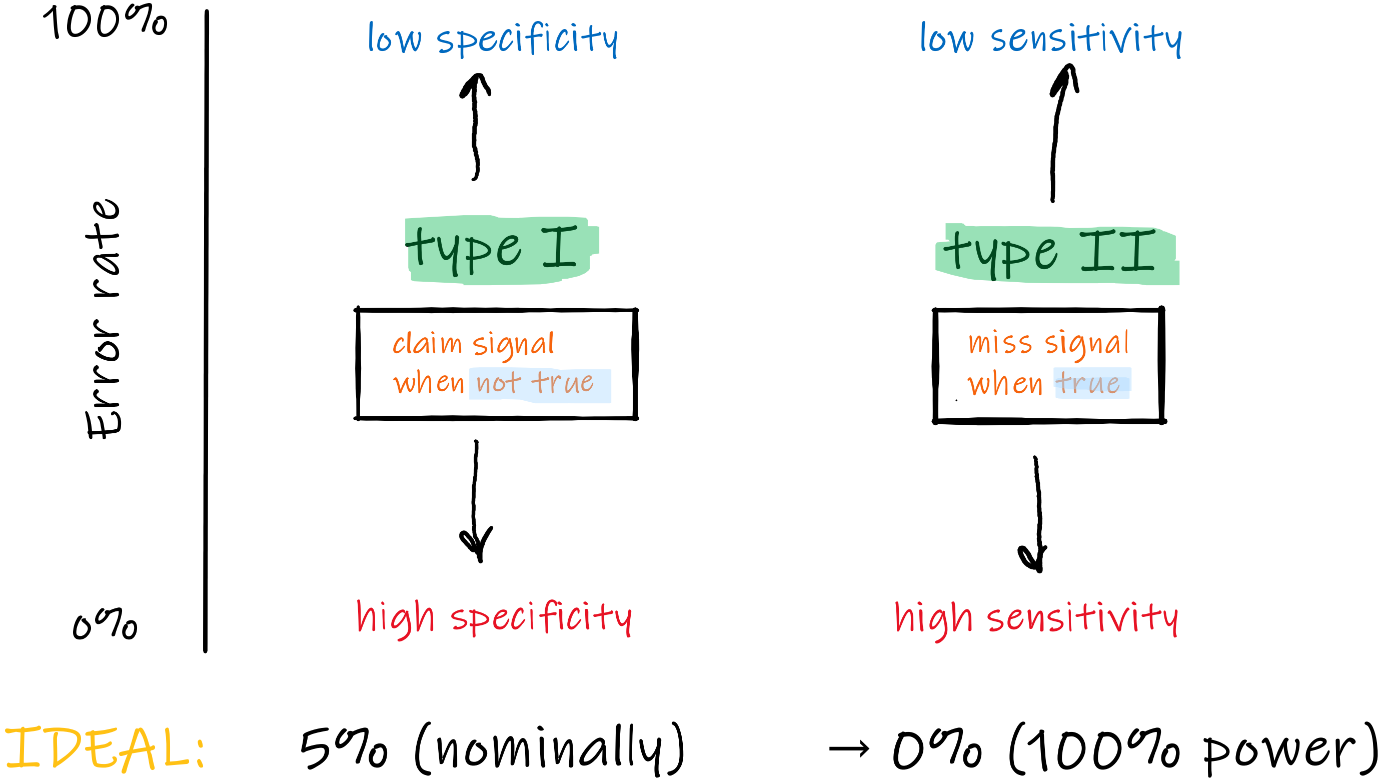
Time at risk and test-statistics

- A ***time-at-risk*** when safety events can occur:
  - 0-1 days, ***1-28 days*** and 1-42 days after vaccination
  - Dose definition (first, second, both)

- The ***test-statistic*** to estimate:
  - Effect-size estimation (incidence rate ratio, hazard ratio or odds ratio)
  - Log-likelihood ratio (common in vaccine surveillance, allows for corrections for multiple testing over time via ***MaxSPRT***)
  - With and without ***empirical calibration*** (to control for systematic error)

Method performance metrics

- A ***decision rule*** to classify signals from non-signals
  - Bias / variance (particularly of the ***residual systematic error***)
  - Type 1 error rate
  - Type 2 error rate
  - ***Timeliness*** to achieve power



Real-world evidence with 117M estimates

Exposures of interest:

- H1N1pdm (`09-`10)
- Seasonal influenza (Fluvirin, `17-`18)
- Seasonal influenza (Fluzone, `17-`18)
- Seasonal influenza (all, `17-`18)
- Zoster (2018, 2 doses)
- HPV (2018, 2 doses)

Data sources:

- CCAE
- MDCR
- MDCD
- Optum EHR

Negative control outcomes (93):

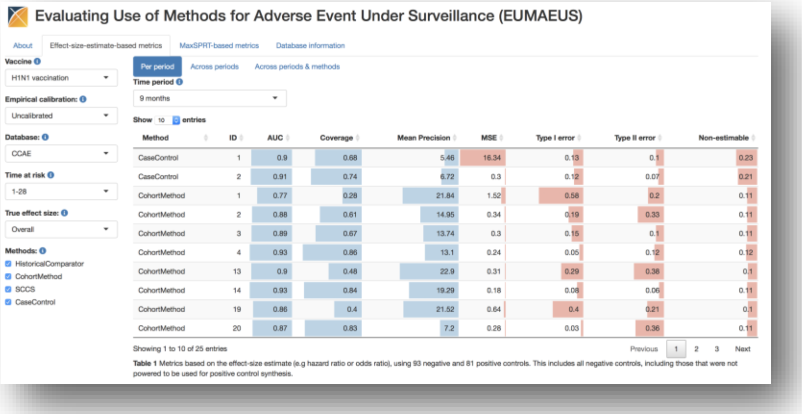
- Not related to any of these vaccines
- Similar prevalence and %-inpatient diagnoses (severity) to AESI
- Clinical expert review

Positive control outcomes:

- Imputed from negative controls
- Known effect sizes (1.5, 2, 4 x)

Open Science: pre-specified and registered protocol, open-source analytic code, public access to all results

– https://ohdsi-studies.github.io/Eumaeus/Protocol.html
– https://github.com/ohdsi-studies/Eumaeus/
– https://data.ohdsi.org/Eumaeus/



Evaluating Use of Methods for Adverse Event Under Surveillance (EUMAEUS)

Vaccine: H1N1 vaccination

Empirical calibration: Uncalibrated

Database: CCAE

Time at risk: 1-28

True effect size: Overall

Methods:
☑ HistoricalComparator
☑ CohortMethod
☑ SCCS
☑ CaseControl

Time period: 9 months

Show 10 entries

Method	ID	AUC	Coverage	Mean Precision	MSE	Type I error	Type II error	Non-estimable
CaseControl	1	0.9	0.68	5.46	16.34	0.13	0.1	0.23
CaseControl	2	0.91	0.74	6.72	0.3	0.12	0.07	0.21
CohortMethod	1	0.77	0.28	21.84	1.52	0.58	0.2	0.11
CohortMethod	2	0.88	0.61	14.95	0.34	0.19	0.33	0.11
CohortMethod	3	0.89	0.67	13.74	0.3	0.15	0.1	0.11
CohortMethod	4	0.93	0.86	13.1	0.24	0.05	0.12	0.12
CohortMethod	13	0.9	0.48	22.9	0.31	0.29	0.38	0.1
CohortMethod	14	0.93	0.84	19.29	0.18	0.08	0.06	0.11
CohortMethod	19	0.86	0.4	21.52	0.64	0.4	0.21	0.1
CohortMethod	20	0.87	0.83	7.2	0.28	0.03	0.36	0.11

Showing 1 to 10 of 25 entries

Table 1 Metrics based on the effect-size estimate (e.g hazard ratio or odds ratio), using 93 negative and 81 positive controls. This includes all negative controls, including those that were not powered to be used for positive control synthesis.

Prelude to the results

- Which methods are ***least bias*** in the real-world?
  - Effect of counterfactual anchoring
  - Effect of confounding adjustment
- What is the ***trade-off*** to achieve, say, 50% power?
- Should we ***combine multiple designs*** (signal generation / evaluation) to improve performance?
- Is ***sequential testing (α-spending) correction*** a panacea?
- Do ***2nd doses*** influence method choice?

Bias, precision and timeliness of historical rate comparison methods

Xintong Li

on behalf of the EUMAEUS task force

Recall the advantages of historical comparator design:

- Greater statistical power
- Improved timeliness

Especially useful at early stage after vaccine introduction

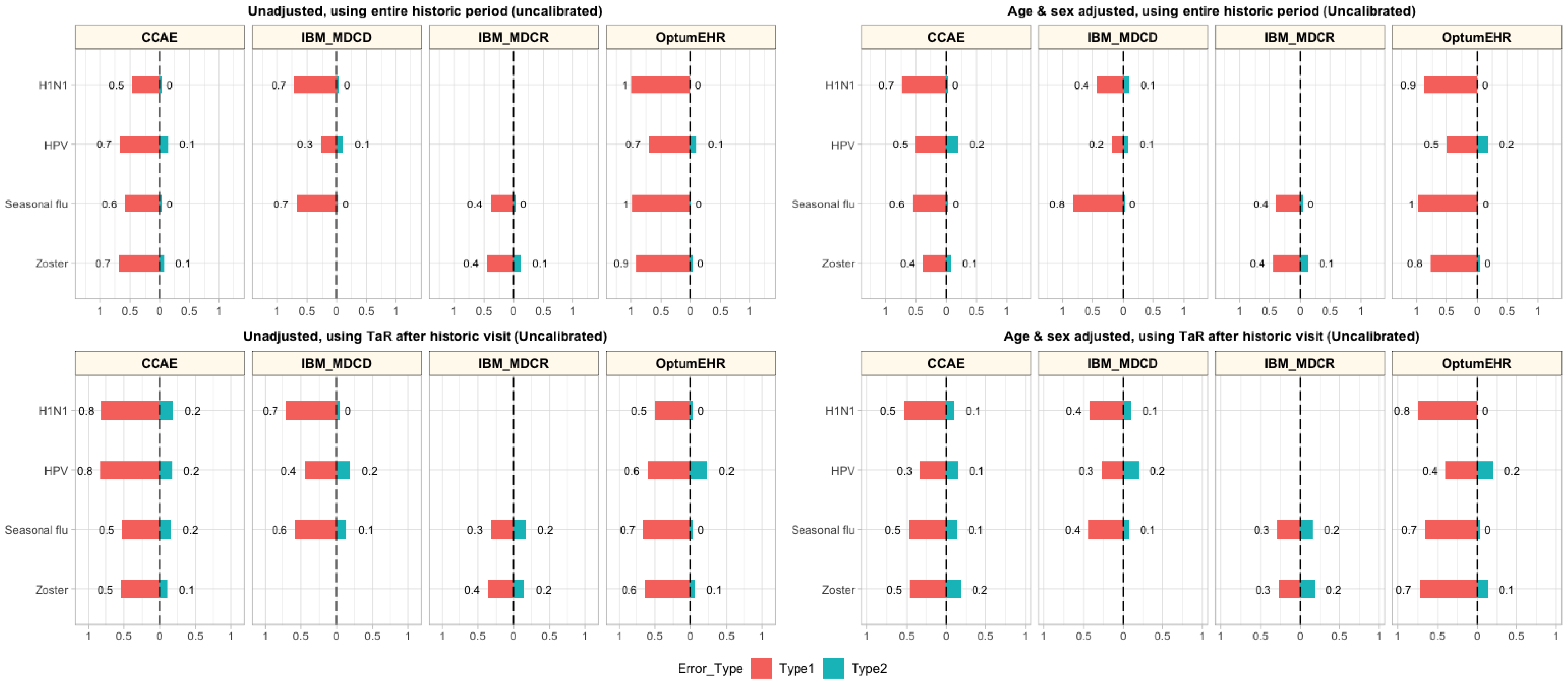
Historical comparator is from:

- literature
- **within same database / population (best-case scenario)**
- others

Choice of design

Population	Time-at-risk	Calibration
Unadjusted	Entire year	Yes
Adjusted for age and sex	Relative to outpatient visit	No

Historical comparison in general: Sensitive but not specific



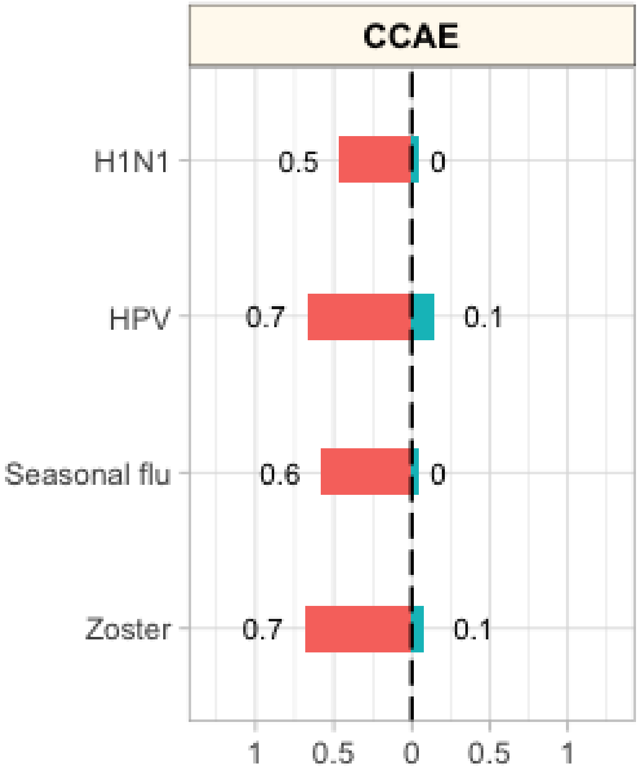
Unadjusted, using entire historic period (uncalibrated)

Age & sex adjusted, using entire historic period (Uncalibrated)

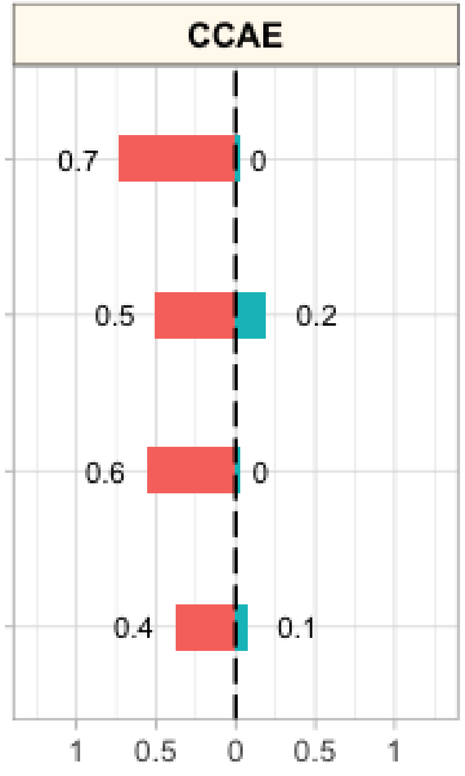
Unadjusted, using TaR after historic visit (Uncalibrated)

Age & sex adjusted, using TaR after historic visit (Uncalibrated)

Sensitive but not specific



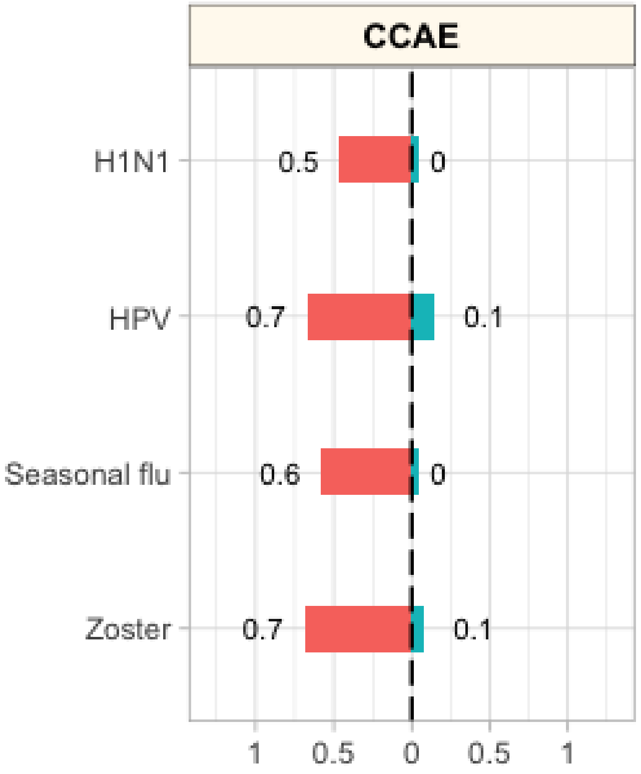
Unadjusted,
entire historical period



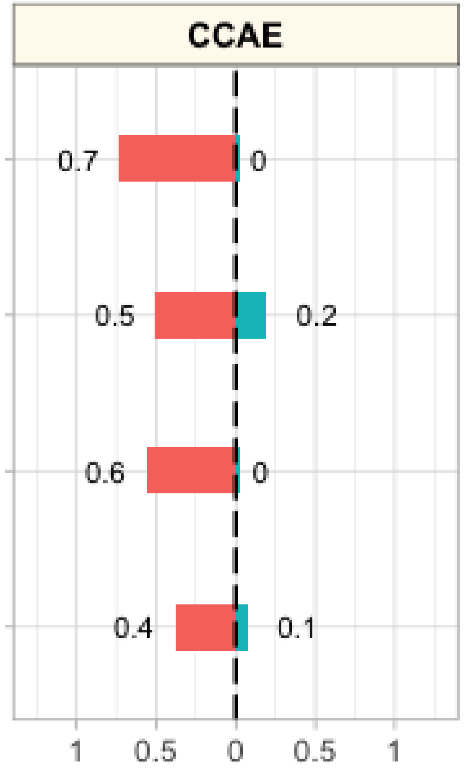
Age and sex adjusted,
entire historical period

Adjust for age and sex reduced type 1 error.

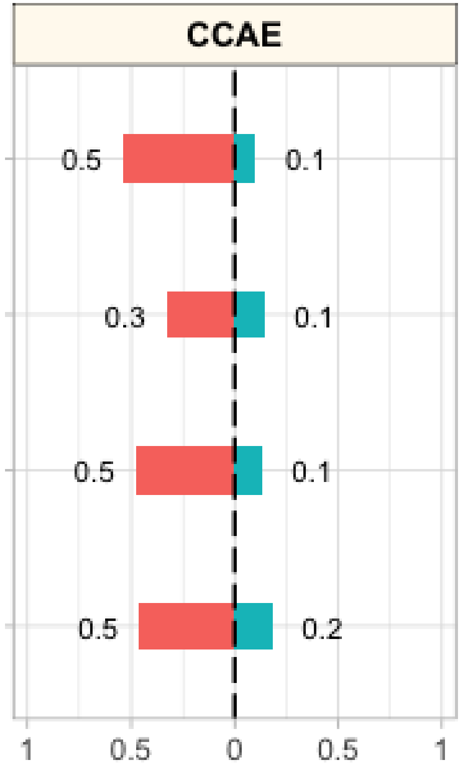
Sensitive but not specific



Unadjusted,
entire historical period



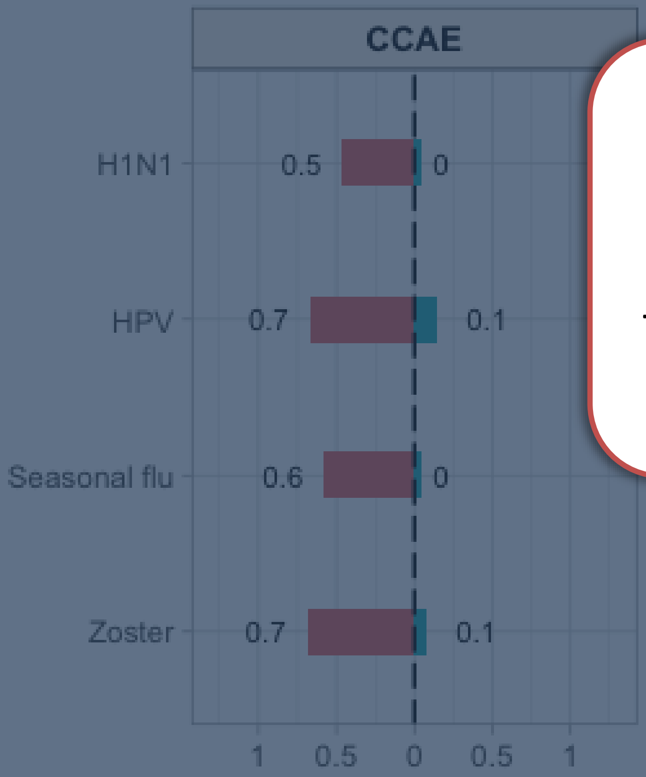
Age and sex adjusted,
entire historical period



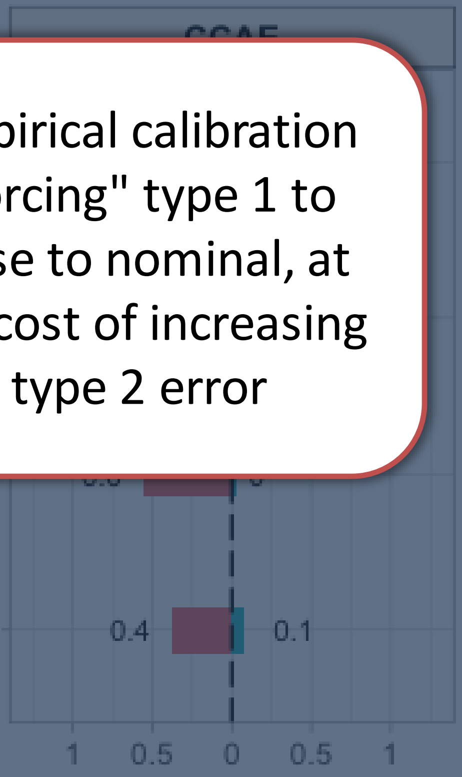
Age and sex adjusted,
Time-at-risk after
historic visit

After adjusting for age and sex, anchoring on visit further reduce type 1 error.

Empirical calibration

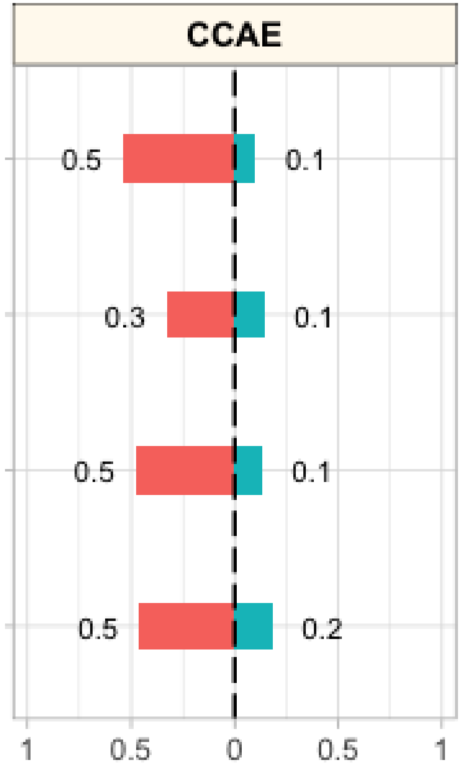


Unadjusted,
entire historical period

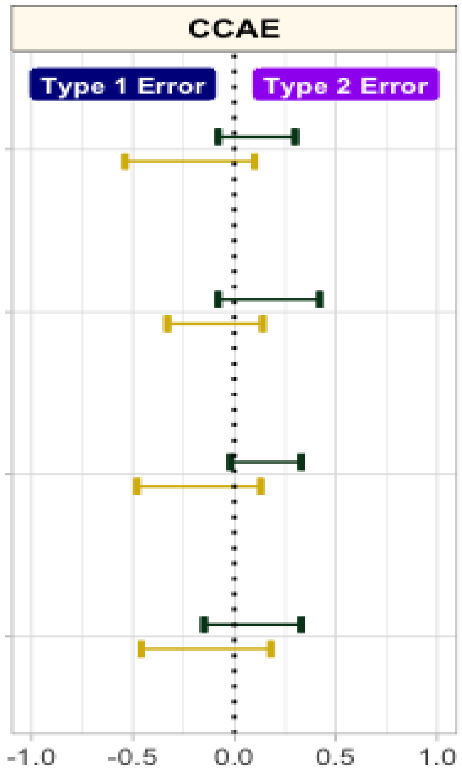


Age and sex adjusted,
entire historical period

Empirical calibration "forcing" type 1 to close to nominal, at the cost of increasing type 2 error

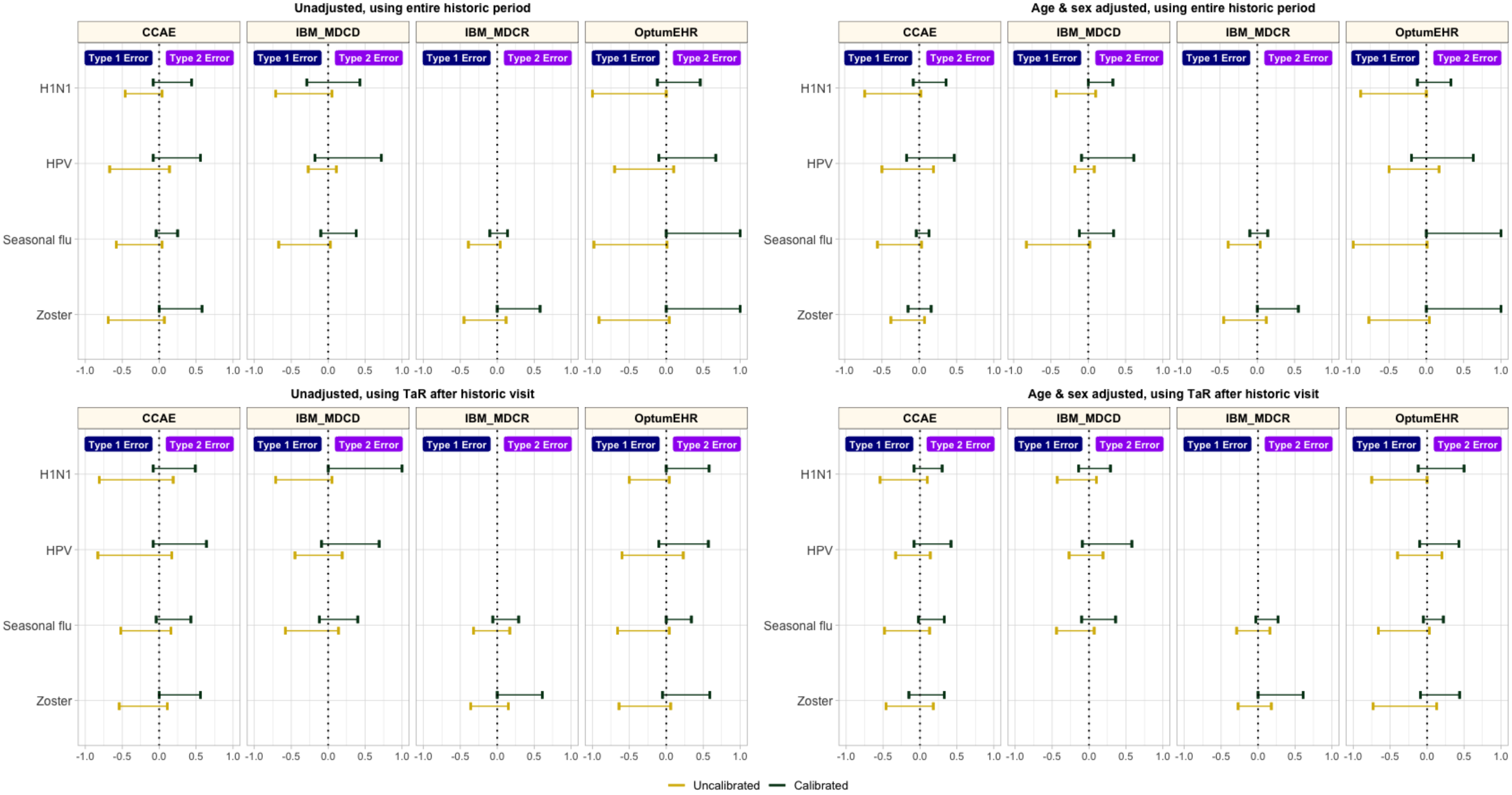


Age and sex adjusted,
Time-at-risk after
historic visit

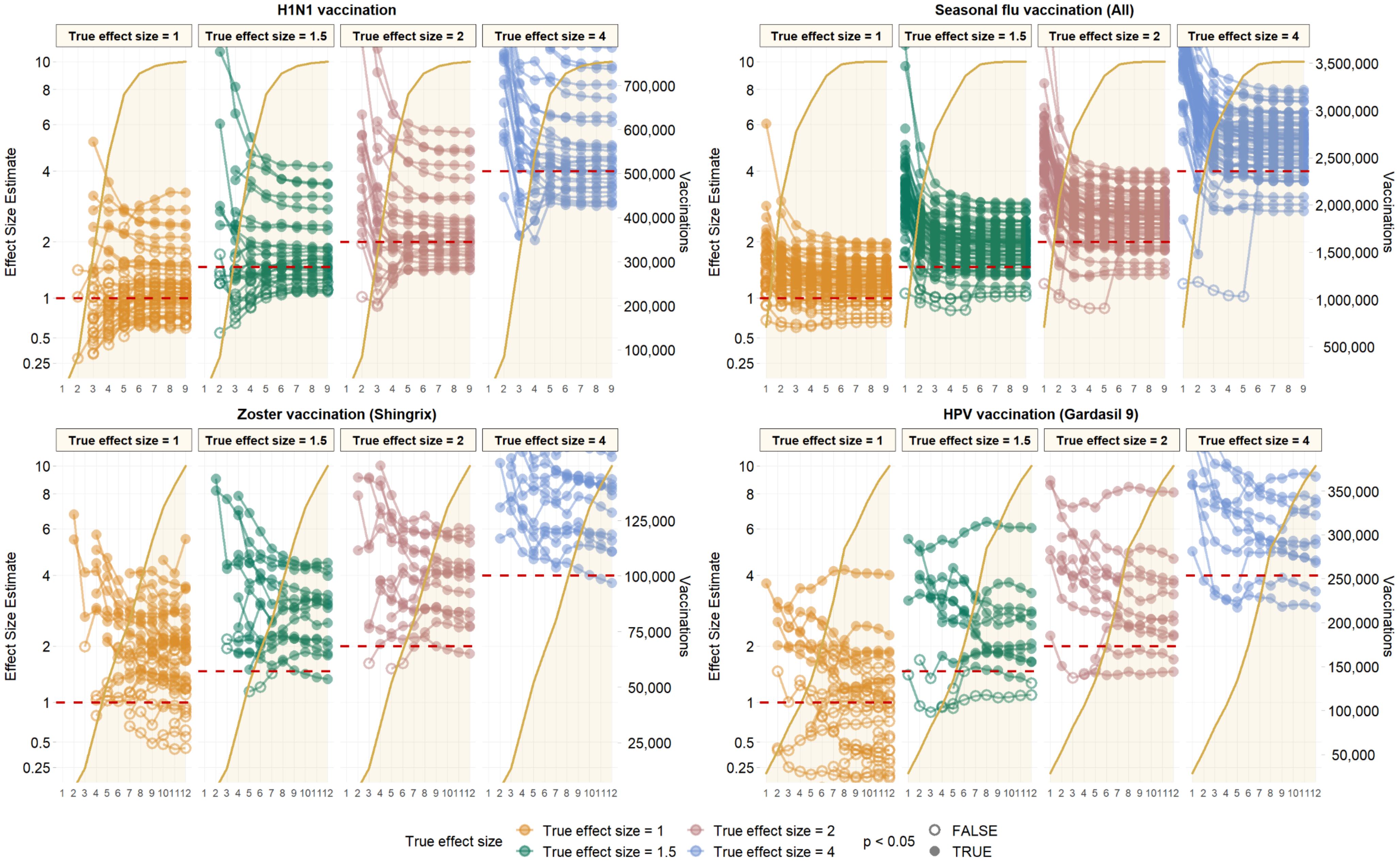


Age and sex adjusted,
Time-at-risk after
historic visit

Empirical calibration: reduce type 1, increase type 2



Higher and faster uptake, earlier detection



Database

CCAE

Analysis

Adjusted for age and sex, no anchoring

Calibration

No

Conclusion

- Sensitive but not specific: overestimate risks
- Age-sex adjustment reduce false positive
- Anchoring on visit reduce false positive
- Empirical calibration: forced type 1 error back to normal, at the cost of increasing type 2 error.
- For vaccine with high uptake speed: can detect earlier, stabilized estimation.

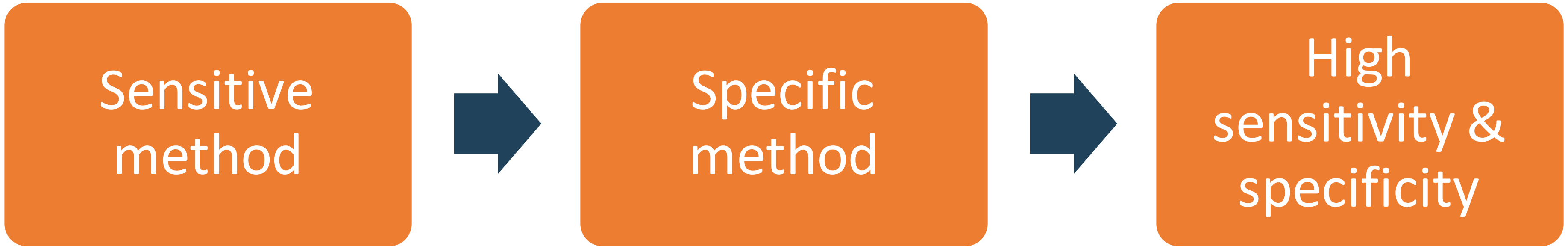
OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Combining Methods in a Safety Surveillance System

Faaizah Arshad

on behalf of the EUMAEUS task force

Introduction



Sensitive method

Specific method

High sensitivity & specificity

- HIV testing
  - – Two part test: 1) highly sensitive (few false negatives); 2) highly specific (eliminate false positives)

Methods

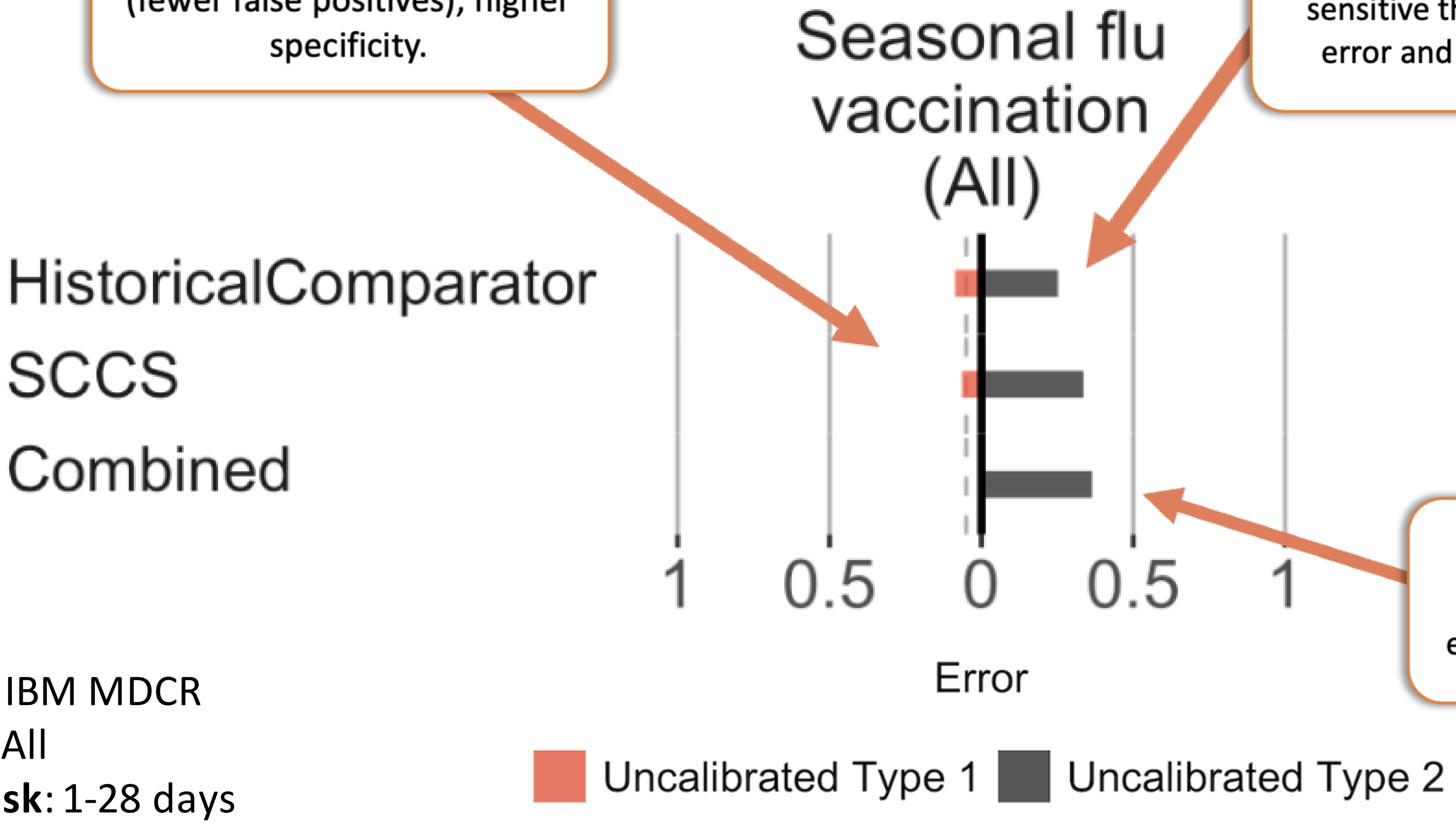
- We hypothesized that sequentially combining methods might be desirable for population-level COVID-19 vaccine safety surveillance.

- Method 1: historical comparator (sensitive / cheap)
- Method 2: self-controlled case series (specific)
- Combined: Method 1 → Method 2

Uncalibrated type I and II errors for all outcomes

SCCS shows lower type I error (fewer false positives); higher specificity.

Historical comparator remains more sensitive than SCCS. Lower type II error and fewer false negatives.

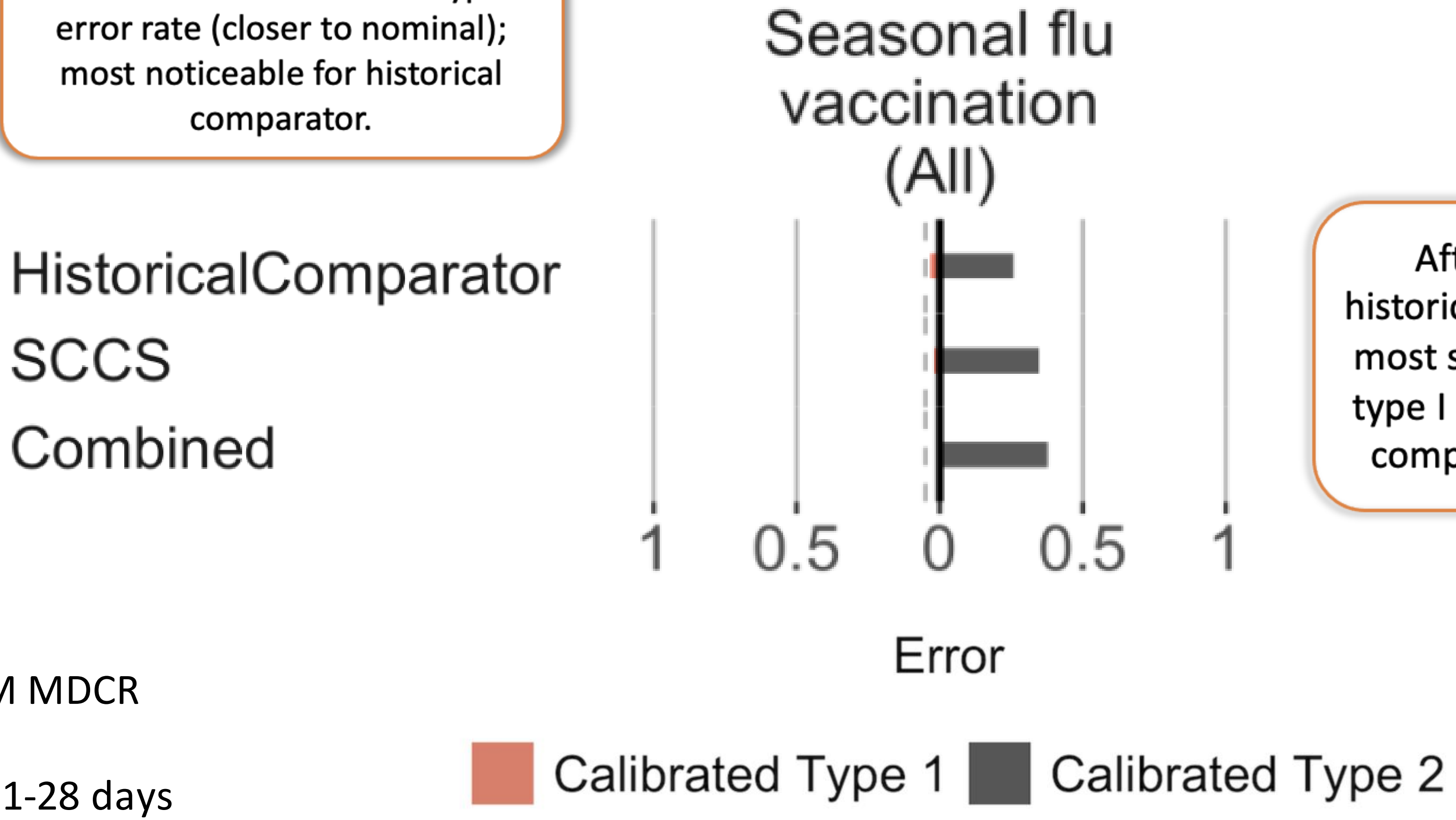


Combined method shows low type I error and greater type II error (specific but not sensitive).

Database: IBM MDCR
Outcome: All
Time-at-Risk: 1-28 days

Calibrated type I and II errors for all outcomes

Calibration tries to fix the type I error rate (closer to nominal); most noticeable for historical comparator.



After calibration, historical comparator still most sensitive. Reduced type I error for historical comparator and SCCS.

Database: IBM MDCR
Outcome: All
Time-at-Risk: 1-28 days

Conclusion

- Reject hypothesis.
- Sequentially combining sensitive and specific methods does not improve performance over using a single method.
- Future vaccine monitoring should consider the sequence of methods used to ensure accurate signal detection.

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Estimation for Two-Dose Vaccines

Ty Stanford

on behalf of the EUMAEUS task force

Aim:

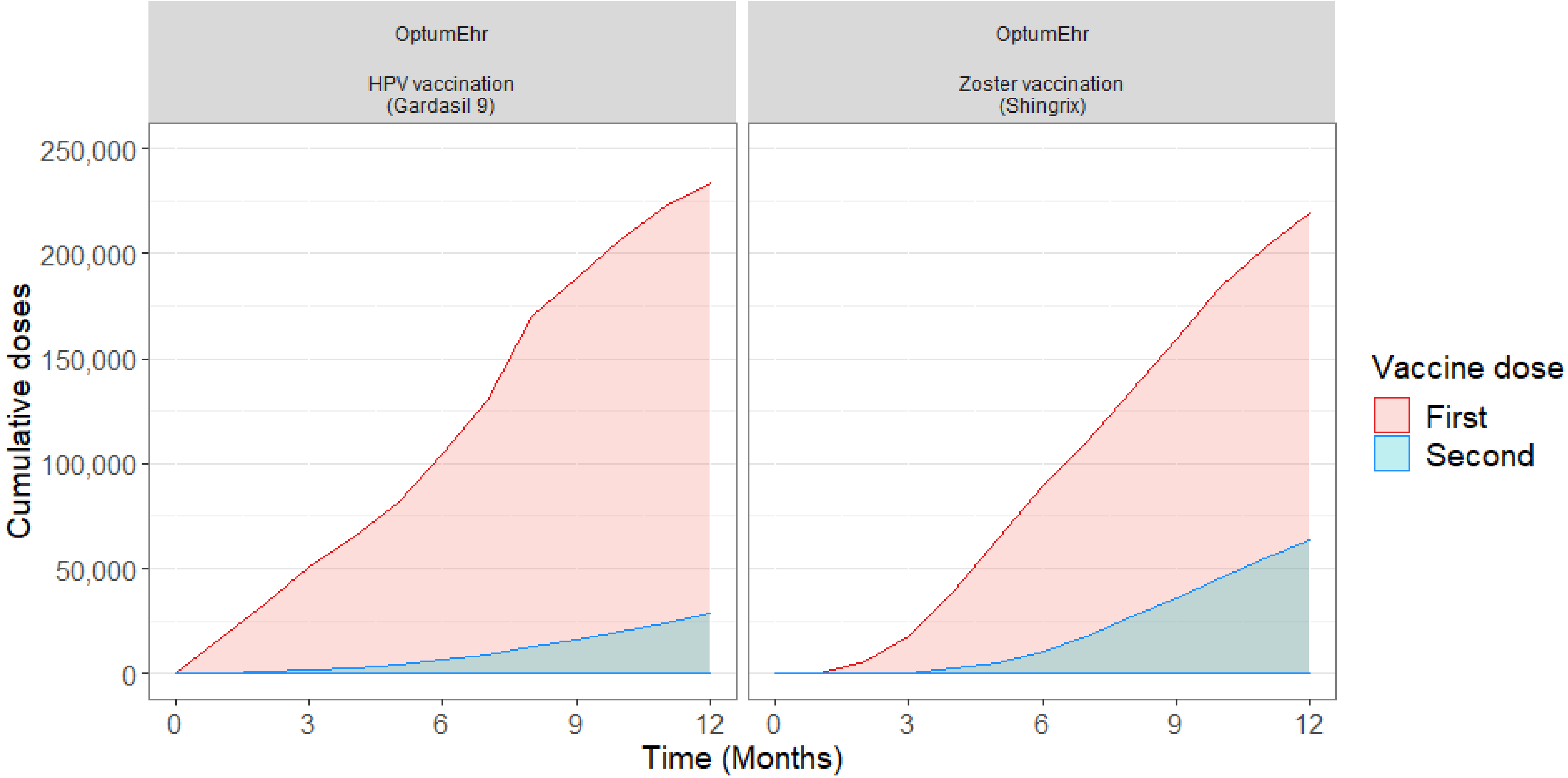
- Does the inclusion of data from the 2nd dose, among vaccines with 2 doses, reduce type II error?

Data:

- This limits EUMAEUS data to

  (CCAE, Optum EHR) x (HPV vaccine, Zoster vaccine) combinations

Dose accumulation



Database	Dose	HPV vaccination (Gardasil 9)	Zoster vaccination (Shingrix)
Optum EHR	First	233985	219665
	Second	28336	63464

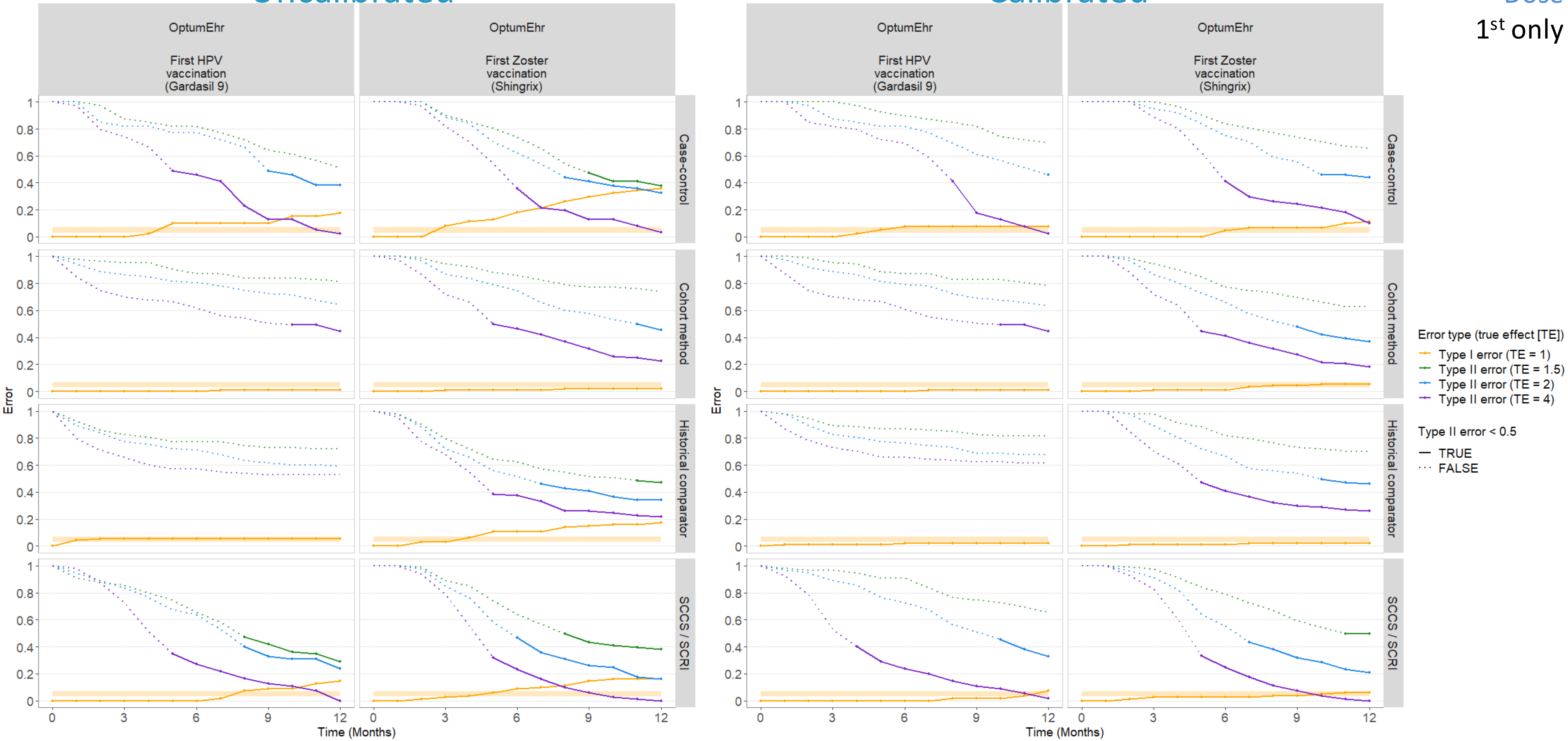
To calibrate or not to calibrate?

Adj. for sequential testing
MaxSPRT

Uncalibrated

Calibrated

Dose
1st only



Adding 2nd dose: Cohort Design

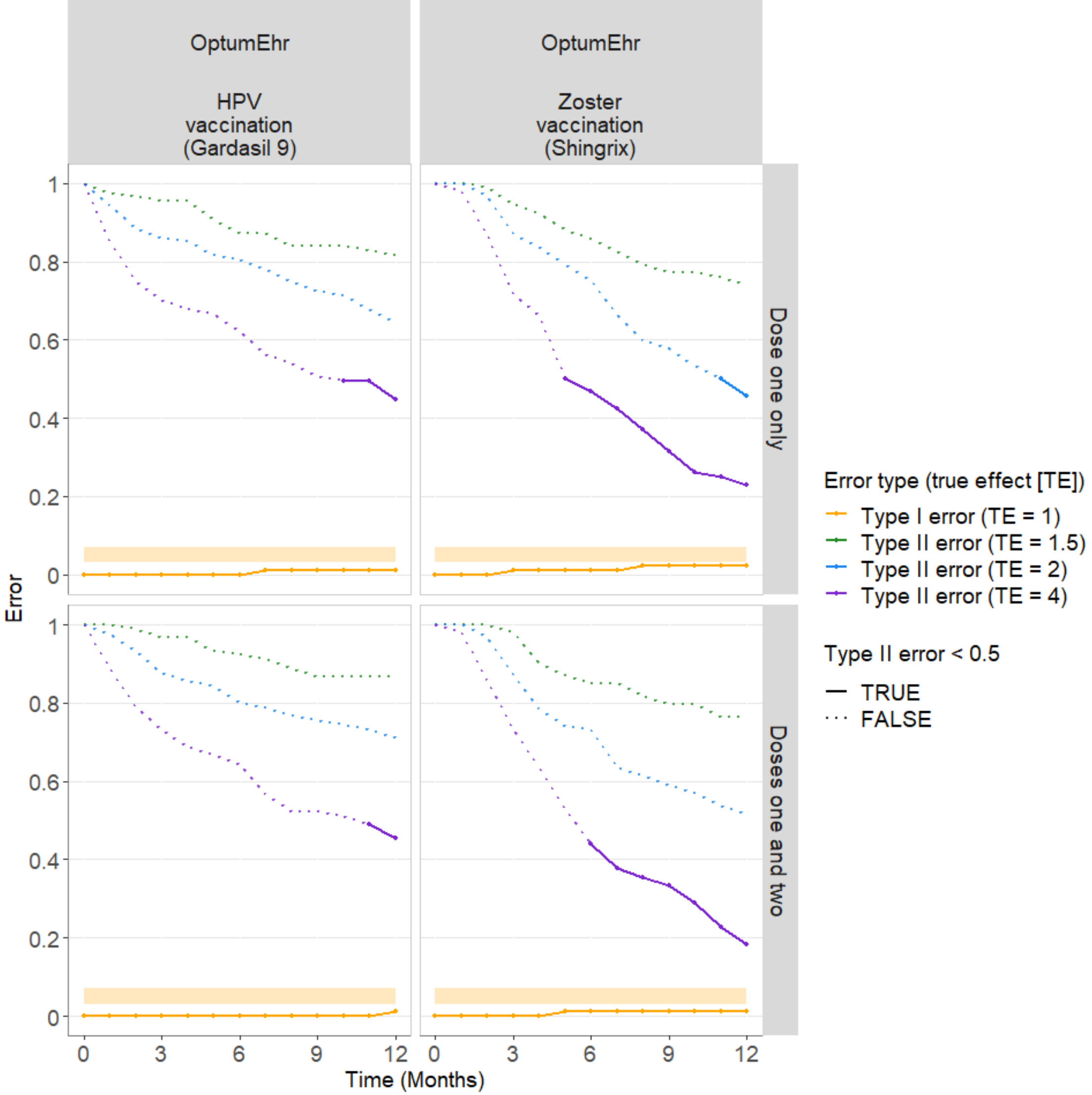
Adj. for sequential testing

MaxSPRT

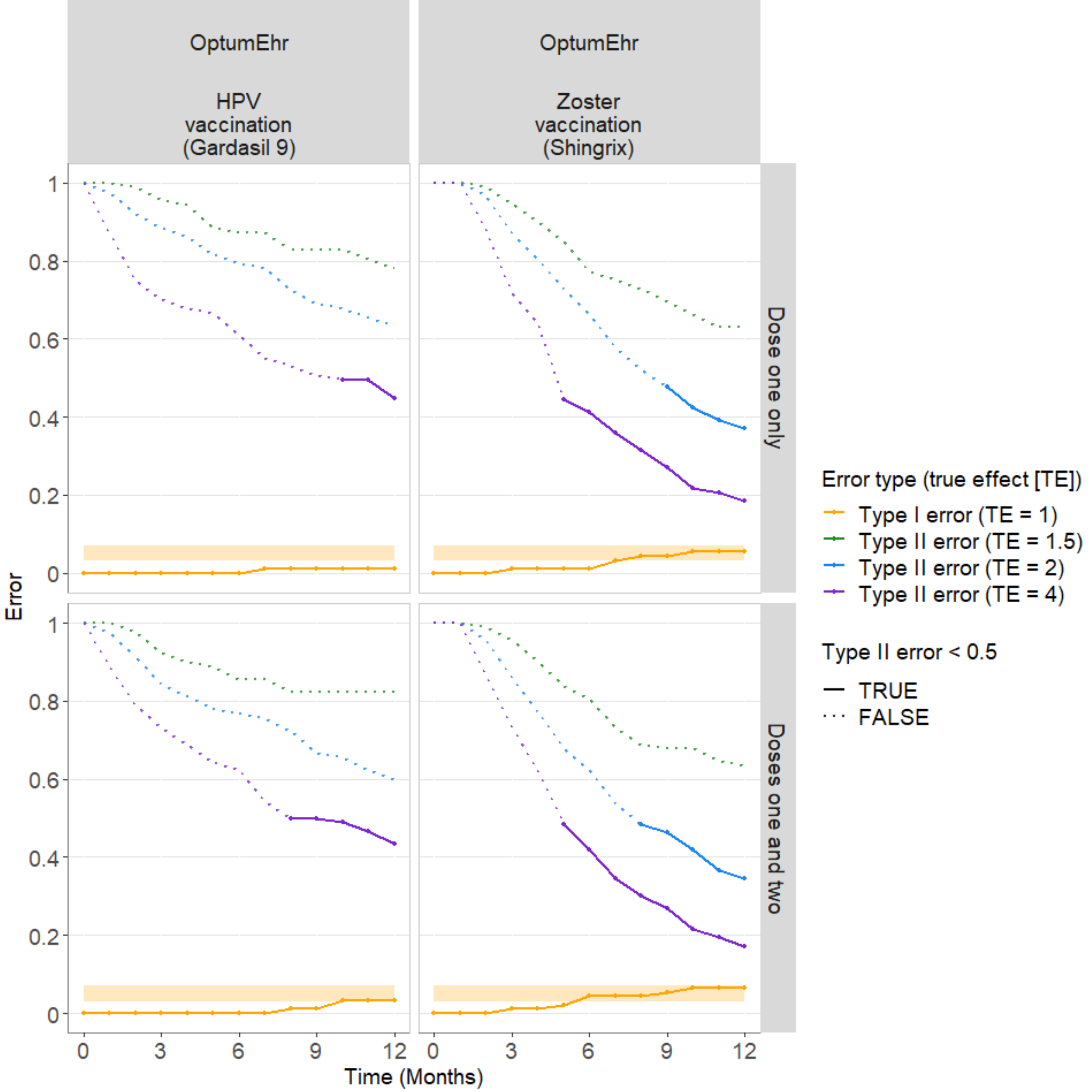
Dose

(1^{st} only) vs (1^{st} & 2^{nd})

Uncalibrated



Calibrated

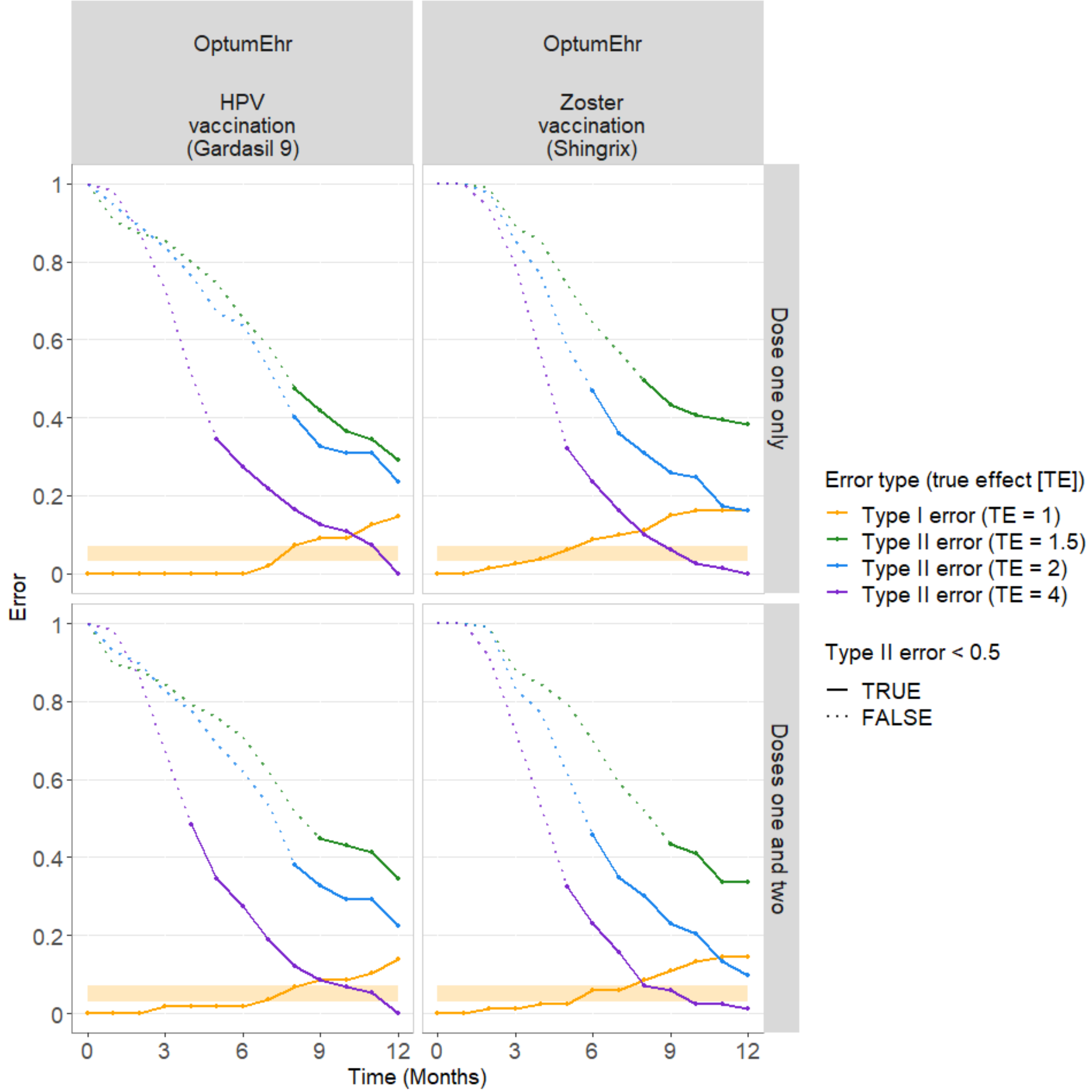


Adding 2nd dose: SCCS

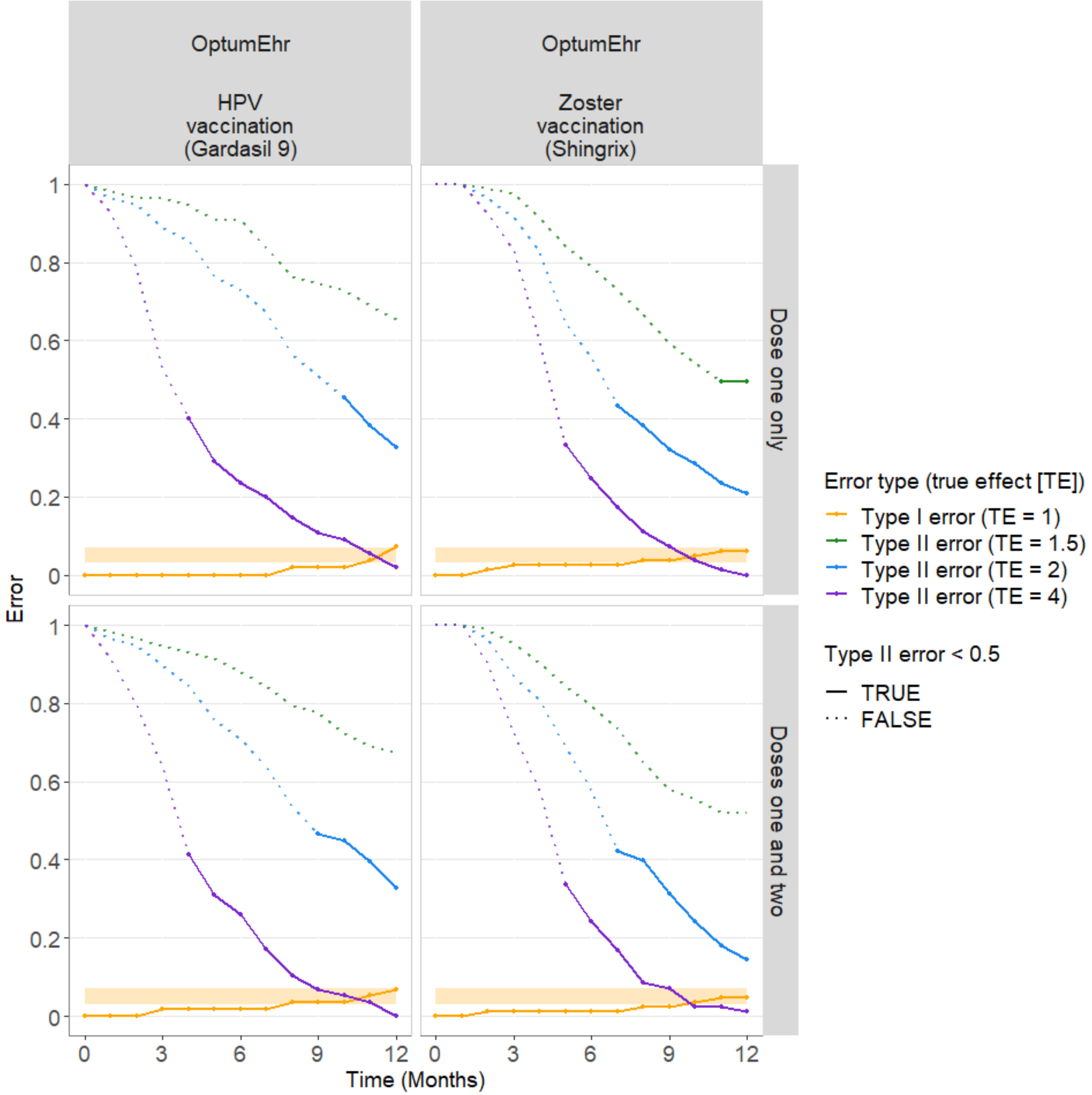
Adj. for sequential testing
MaxSPRT

Dose
(1st only) vs (1st & 2nd)

Uncalibrated



Calibrated



Conclusion

- Inclusion of the 2nd dose can increase the power
  - marginally in this case, likely as a result of a marginal increase in sample size
- The most important factor is *empirical calibration*
  - more data doesn't magically negate issues with specific designs
- Future work to understand the issues better:
  - Larger proportion of 1st doses to also have 2nd doses (with differing rates)
  - Underlying signals (positive controls) to have varying effects after each dose

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

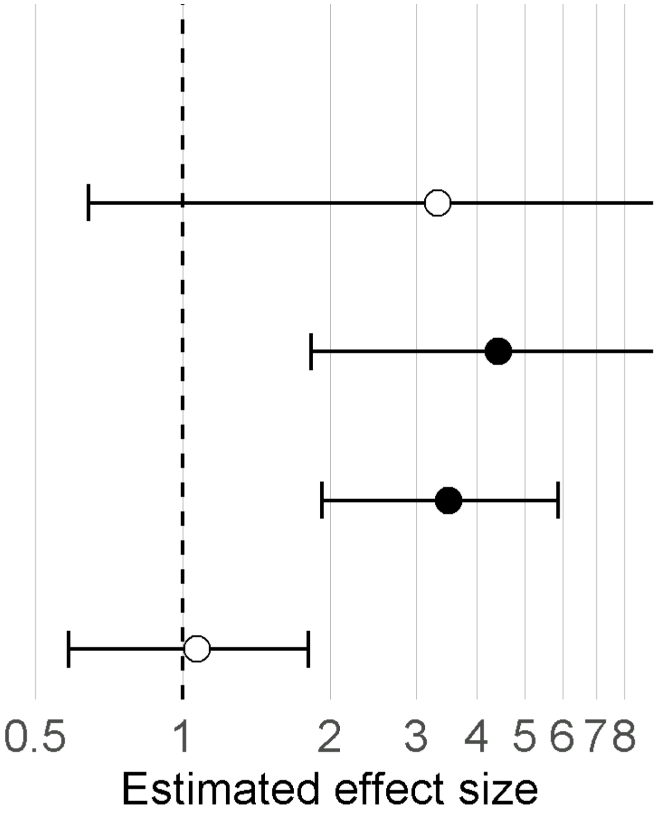
Comparison of performance across methods

Martijn Schuemie

on behalf of the EUMAEUS task force

Same data & question, different methods: different results

Analysis choices	Effect size (95% CI)
Matched case-control	3.33 (0.64-15.44)
PS-weighted cohort method	4.42 (1.83-10.42)
Historical comparator	3.50 (1.92-5.87)
SCCS	1.07 (0.59-1.81)

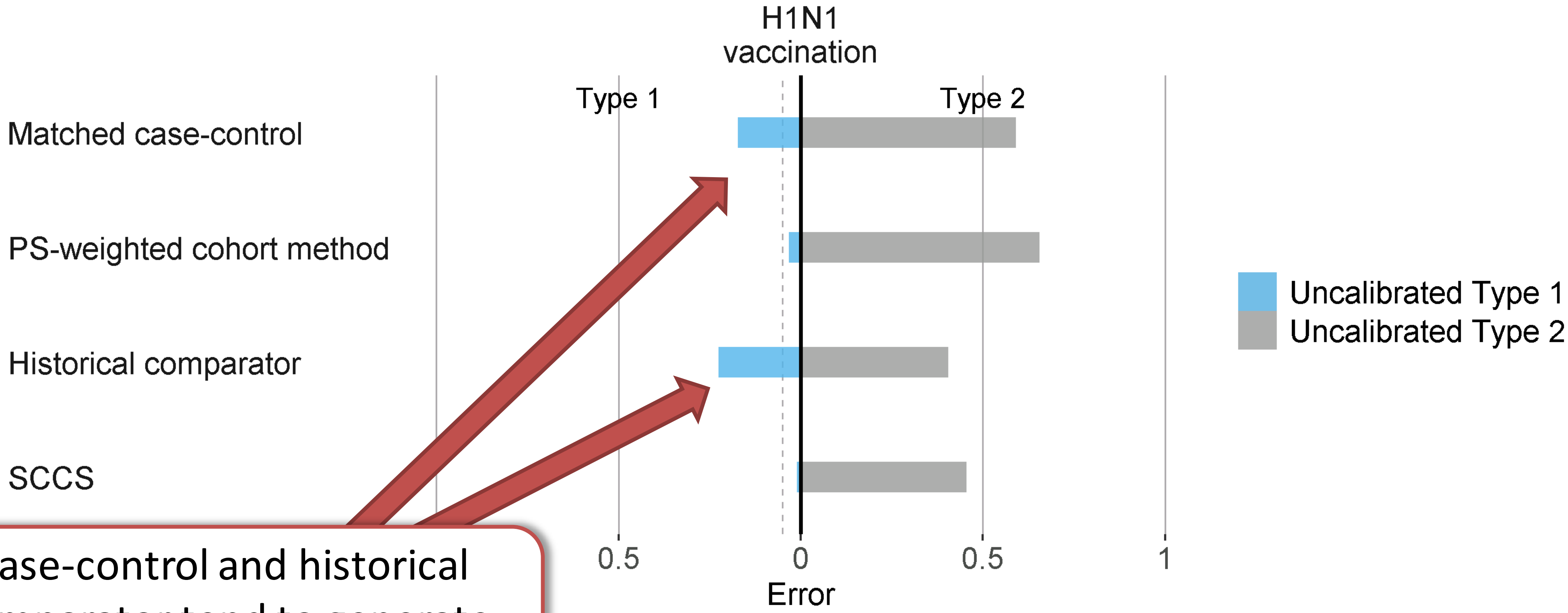


Exposure
H1N1pdm vaccinations

Outcome
Contusion of toe

Database
Optum EHR

Comparing on type 1 and type 2 error

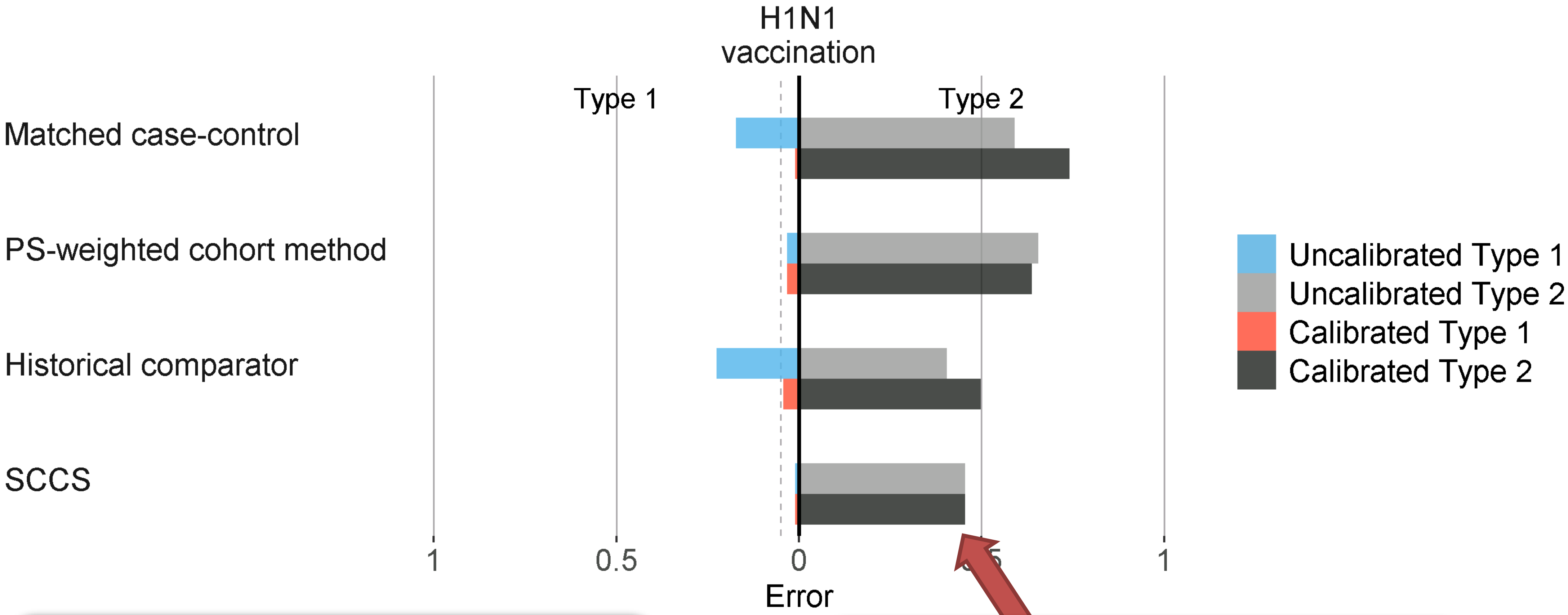


Case-control and historical comparator tend to generate many false positives

Database

Optum EHR

Empirical calibration: restoring type 1



Calibration makes methods comparable.

After calibration (fixed type 1), SCCS has lowest type 2 error

Database
Optum EHR

Adjusting for systematic error and sequential testing

	Type 1 error	
	Historical comparator	SCCS
Uncalibrated, no adjustment for sequential testing	28.0%	4.3%
Uncalibrated, MaxSPRT	18.3%	2.2%
Calibrated, no adjustment for sequential testing	10.8%	5.4%
Calibrated, MaxSPRT	6.5%	4.3%

New!

Exposure

H1N1pdm vaccinations

Outcome

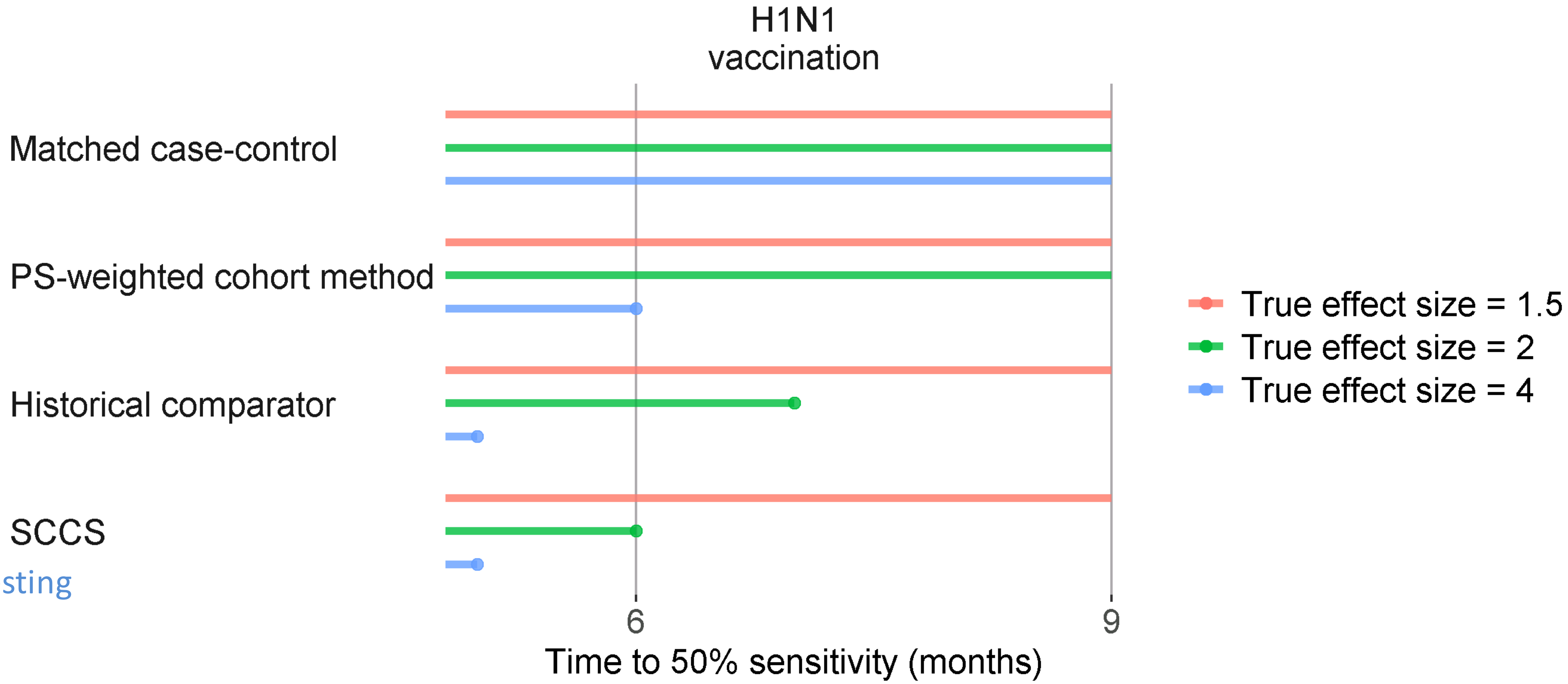
All negative controls

Database

Optum EHR

Adjusting for **systematic error**
has bigger impact than
adjusting for **sequential testing**

Time to 50% sensitivity (after calibration)



Adj. for sequential testing

MaxSPRT

Adj. for systematic error

Empirical calibration

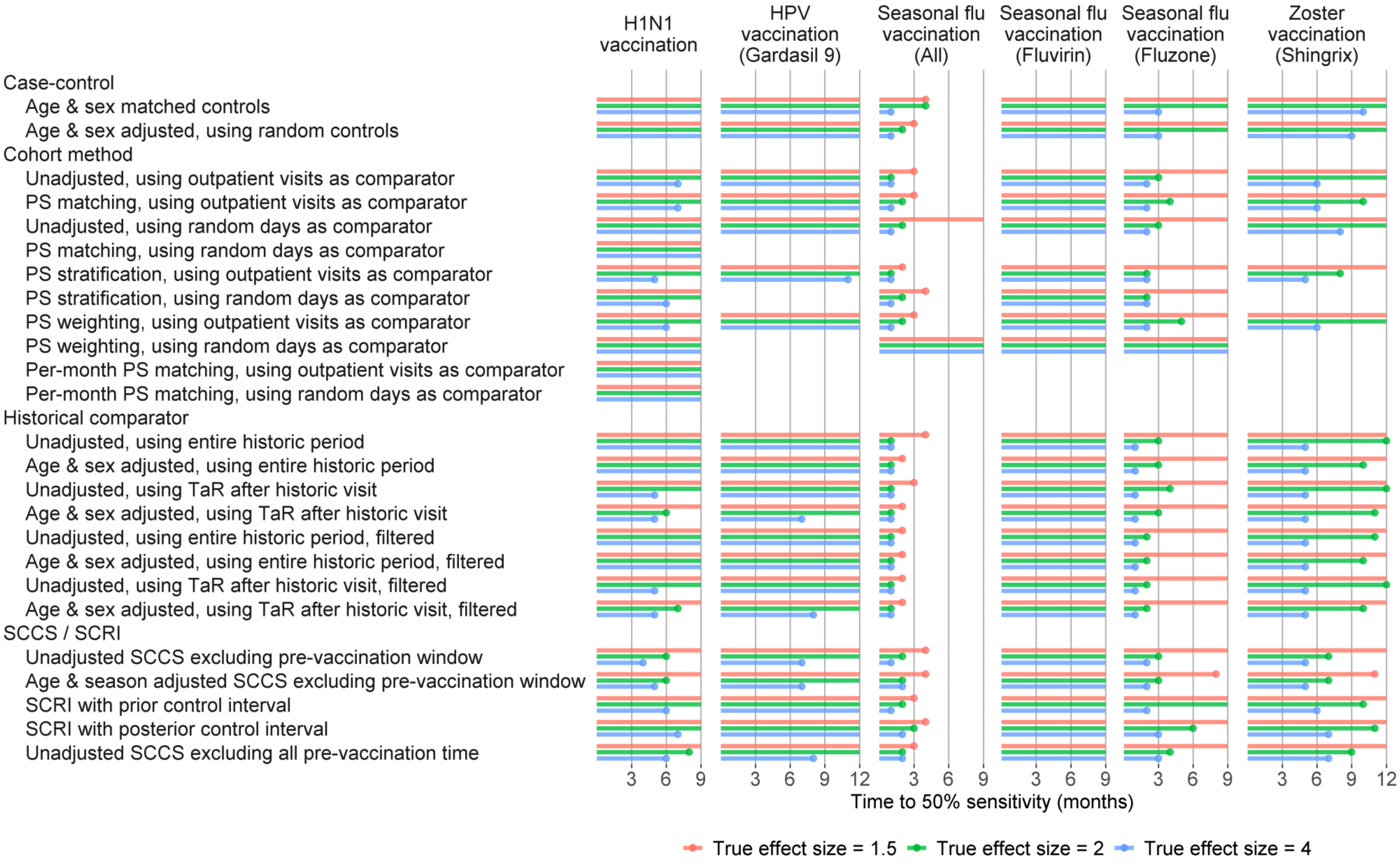
Database

Optum EHR

SCCS has shortest time to detection

All methods struggle to achieve 50% sensitivity for small effects

More or less consistent across methods / outcomes /databases



Adj. for sequential testing

MaxSPRT

Adj. for systematic error

Empirical calibration

Database

Optum EHR

Conclusions

- Many methods show large systematic error / type 1 error
- Empirical calibration can restore type 1 error to nominal, at the cost of increasing type 2 error

  (depending on magnitude of systematic error)
- Empirical calibration often has bigger impact than adjusting for sequential testing

  (should do both)
- After calibration and adj. for sequential testing SCCS seems overall best

  (shortest time to detection)
- No method achieves high sensitivity for small true effect sizes

  (on these data)

Recommendations for a safety surveillance system

Martijn Schuemie

on behalf of the EUMAEUS task force

Recommendations

- Many methods (e.g. case-control & historical comparator) have positive bias, causing many false positives (high type 1 error)
  - Include negative controls and use empirical calibration
  - Include self-controlled designs
  - Always use confounding adjustment
  - Carefully consider anchoring of counterfactual
- Detecting more than half of true adverse effects may require accepting more false positives (e.g. using calibrated $p < 0.10$)
- Combining multiple designs likely doesn’t improve performance
  - Do not distinguish between ‘signal generation' and 'signal evaluation'
- Second dose often underpowered to contribute to evidence