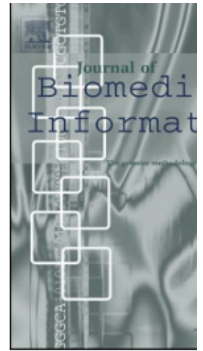




Contents lists available at [ScienceDirect](https://www.sciencedirect.com)




Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Evaluating the Bias, type I error and statistical power of the prior Knowledge-Guided integrated likelihood estimation (PIE) for bias reduction in EHR based association studies

Naimin Jing^{a,1}, Yiwen Lu^{b,c} , Jiayi Tong^{a,b,i}, James Weaver^d , Patrick Ryan^d, Hua Xu^e,
Yong Chen^{a,b,c,f,g,h,*} 

^a Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

^b Center for Health AI and Synthesis of Evidence (CHASE), Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

^c The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA

^d Janssen Research & Development, Titusville, NJ, USA

^e Department of Biomedical Informatics and Data Science, Yale University, New Haven, CT, USA

^f Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA

^g Leonard Davis Institute of Health Economics, Philadelphia, PA, USA

^h Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA

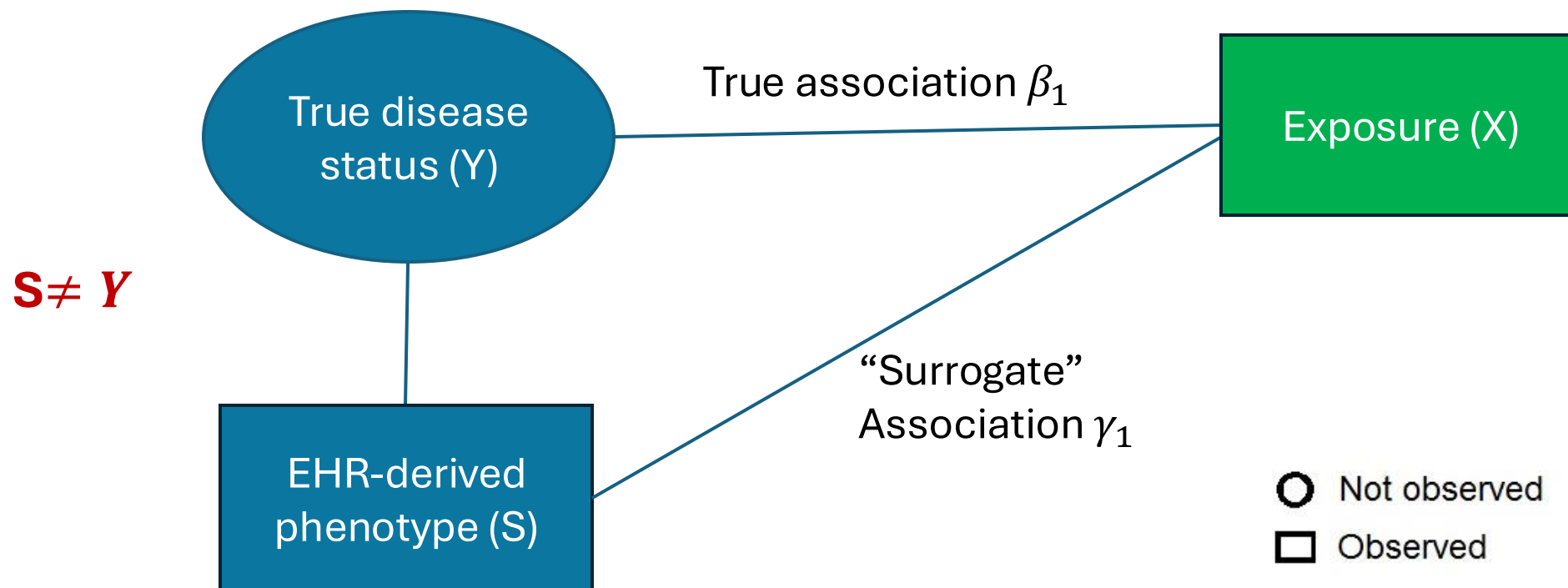
ⁱ Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA



April 1st, 2025

Estimation bias due to phenotyping error

Real-world EHR data often suffer from phenotyping error due to imperfect phenotyping algorithms.



$$\text{logit}\{\Pr(Y_i = 1)\} = \beta_0 + \beta_1 x_i$$

$$\text{logit}\{\Pr(S_i = 1)\} = \gamma_0 + \gamma_1 x_i \text{ (naïve method)}$$

Estimating the true association

$$\Pr(S_i = 1) = (1 - \alpha_0) + (\alpha_0 + \alpha_1 - 1)\text{expit}(\beta_0 + \beta_1 x_i), \text{expit}(t) = \frac{e^t}{1+e^t}.$$

$\alpha_0 = \Pr(S_i = 0|Y_i = 0)$: Specificity of the phenotyping algorithm

$\alpha_1 = \Pr(S_i = 1|Y_i = 1)$: Sensitivity of the phenotyping algorithm

- With known α_0 and α_1 , an unbiased estimator of β_1 can be achieved by maximum likelihood estimation.

$$L(\beta_0, \beta_1, \alpha_0, \alpha_1) = \prod_{i=1}^n \Pr(S_i = 1)^{S_i} (1 - \Pr(S_i = 1))^{1-S_i}$$

- However, it can be hard to determine the correct α_0 and α_1 .

Prior-knowledge-guided Integrated-likelihood Estimation (PIE) method

PIE “average” over a range of possible values by adopting integrated likelihood.

- Maximize the integrated likelihood

$$L_I(\beta_0, \beta_1) = \int \int L(\beta_0, \beta_1, \alpha_0, \alpha_1) \pi(\alpha_0, \alpha_1) d\alpha_0 d\alpha_1$$

- $\pi(\alpha_0, \alpha_1)$ is a given prior distribution
- Requires **specifying only a prior distribution of α_0 and α_1** instead of the value of α_0 and α_1 .

Huang, J., Duan, R., Hubbard, R.A., Wu, Y., Moore, J.H., Xu, H. and Chen, Y., 2018. PIE: A prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *Journal of the American Medical Informatics Association*, 25(3), pp.345-352.

Evaluating PIE's performance

Evaluation questions

How well does PIE perform under a wide spectrum of operating characteristics of phenotyping algorithms under **real-world scenarios**?

From a hypothesis testing point of view, does PIE improve **type I error and statistical power** relative to the naïve method?

How does the **choice of prior distribution** impact the performance of PIE?

Methods

Evaluating PIE on simulated data that are generated under diverse outcome prevalence and association effect sizes, mimicking the real-world setting.

Evaluating PIE on synthetic positive controls of COVID-19 infection constructed based on known negative controls.

Simulation study

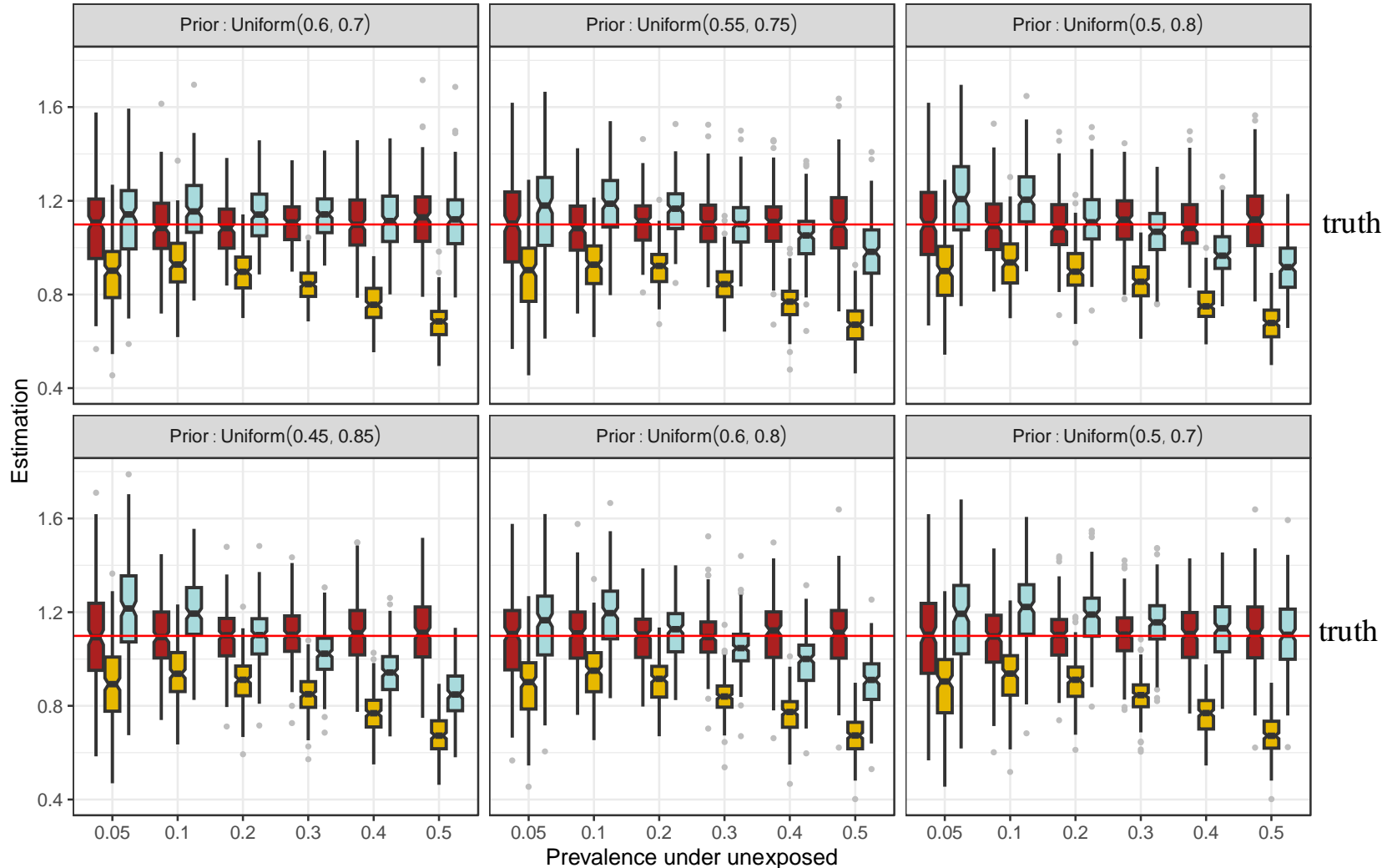
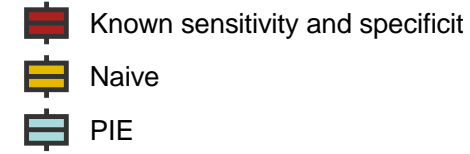
Evaluating the bias, type I error and power of PIE

- The exposure x : Bernoulli distribution with a mean of 30%.
- The prevalence of outcome Y under unexposed (determined by β_0) varies from 5% to 50%
- Effect size (β_1): $\log 3$ in bias evaluation, 0 in type I error evaluation, and varies in power evaluation.
- True specificity: 99%; True sensitivity: 65%
- The prior distribution of specificity is fixed as Uniform (0.95, 0.9999).
- The prior distribution of sensitivities varies under different mean and spread. We used uniform distribution, beta distribution and logit normal distribution.
- Methods: PIE with different priors, naïve method, MLE with known sensitivity and specificity

Simulation study - Bias

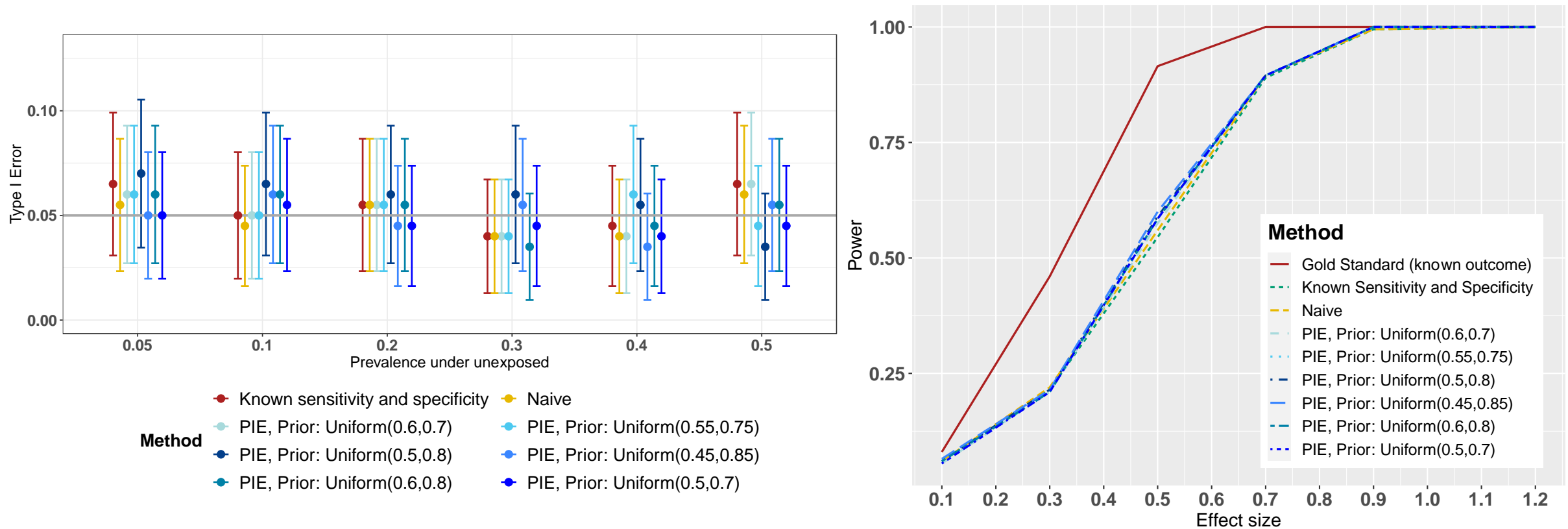
Association estimation from 200 simulated data (N=3000).

Method



- PIE was always closer to the truth compared with the naive method, indicating a **bias reduction**.
- As the prior of sensitivity became more variable (i.e., larger spread), PIE shifted away from the truth.
- The influence of the prior of sensitivity was larger as the prevalence of the response became larger.

Simulation study – Type I Error & Power



- Type I Error: Similar across the methods. No specific pattern.
- Power: Similar across different methods and all smaller than the gold standard method (known outcome).

Real-World-Inspired Evaluation Design

- **Goal:**
 - Evaluate the robustness of PIE under real-world-like conditions.
- **Setting informed by real-world data:**
 - Outcome: COVID-19 infection (binary) emulating institutional phenotype definitions.
 - Predictors: Synthetic *positive controls* created from known *negative controls*.
 - Prevalence, sensitivity, specificity, and missing data patterns informed by real EHRs (e.g., PEDSnet, CHOP studies).
- **Details:**
 - Sample size: 3,000
 - Outcome prevalence: 5% to 50%
 - Sensitivity: 0.65 (low-end of real-world phenotyping)
 - Specificity: 0.99 (typical of real-world EHRs)

Real-World-Inspired Evaluation Design

Negative → Positive Control Construction:

- Based on real-world unassociated predictors: H46-H48, H53-H54, H30-H36, H15-H22.
- Positive controls created by injecting known effect sizes: **1.5 and 4**.
- Maintains realistic covariate distribution and EHR structure.

Approach:

- Applied PIE and naive estimators to bootstrap with 20 re-sampling.
- Priors:
 - Sensitivity \sim Uniform(0.79, 0.95)
 - Specificity \sim Uniform(0.95, 0.9999)

Purpose:

- Mimic real-world evaluation where ground truth is not observable.
- Examine **bias and variability** across effect size magnitudes.

Real-World-Inspired Evaluation Design Result

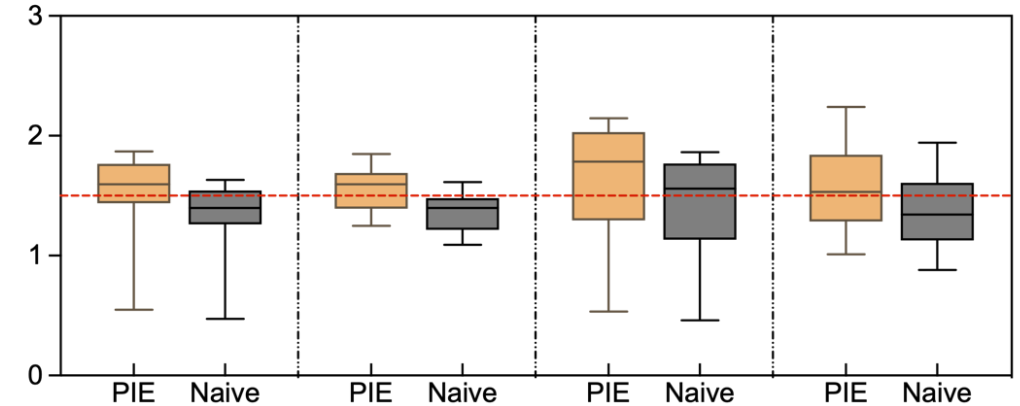
Finding:

- PIE consistently outperformed the naïve estimator, especially at higher effect sizes.
- Naïve method showed systematic attenuation toward the null, growing worse with stronger effects.
- PIE demonstrated robustness even with moderately informative priors.

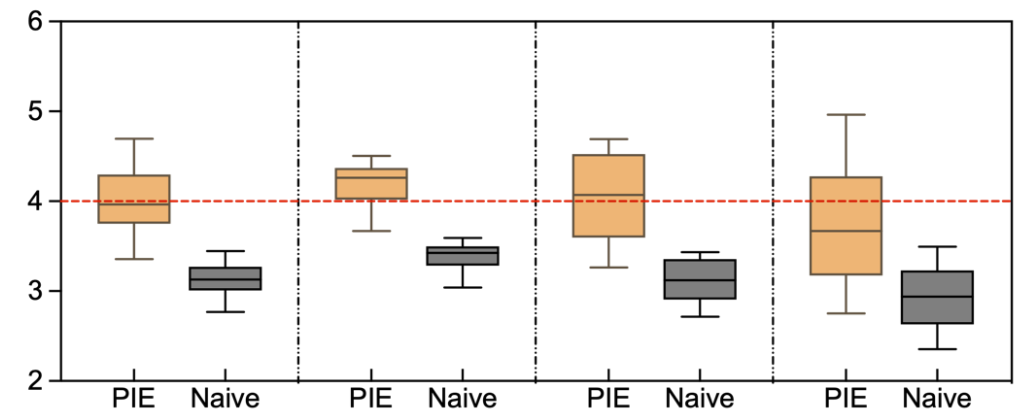
Interpretation:

- Emulated results reflect what would likely occur in real-world EHR studies.
- PIE is most beneficial for estimation, particularly under noisy or uncertain phenotyping conditions.
- Supports future real-world applications, such as trial emulation and vaccine effectiveness research.

(a) Effect Size 1.5



(b) Effect Size 4



Evaluating PIE's performance: Conclusion

Evaluation questions

How well does PIE perform under a wide spectrum of operating characteristics of phenotyping algorithms under **real-world scenarios**?

From a hypothesis testing point of view, does PIE improves **type I error and statistical power** relative to the naïve method?

How **does the choice of prior distribution** impact the performance of PIE?

Conclusion

PIE **effectively mitigates estimation bias** due to phenotyping errors in a wide spectrum of real-world settings, particularly with accurate prior information.

Its main benefit lies in bias reduction rather than hypothesis testing improvement.

The impact of the prior is small for low-prevalence outcomes.

Thank you!