Penn Medicine

Department of Biostatistics, Epidemiology and Informatics

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

The Fine Art of Tolerance: Robustify P-value Calibration in Observational Studies with Partially Valid Negative Control Outcomes

Bingyu Zhang, PhD Candidate, University of Pennsylvania

Advisor: Dr. Yong Chen

CHASE
Center for Health AI
& Synthesis of Evidence

PennCIL
A Computing · Inference · Learning
lab at University of Pennsylvania

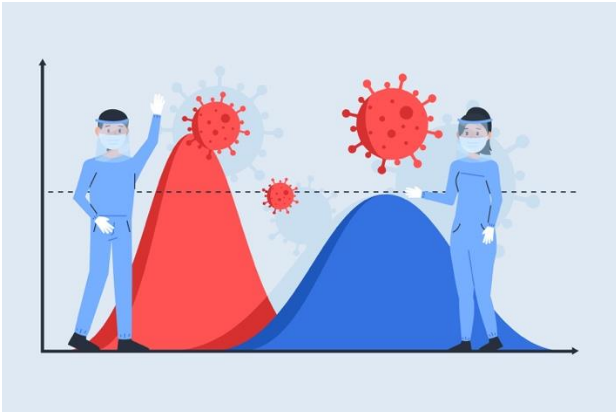
OHDSI Early-Stage Researchers Community Call, November 25, 2025

Joint work with Drs. Dazheng Zhang, Huiyuan Wang, Wenjie Hu, Qiong Wu, Chongliang Luo, Lu Li, Tsai Hor Chan, Yudong Wang, Yuru Zhu, Martijn Schuemie, Patrick Ryan, George Hripcsak, Marc Suchard, and Yong Chen

Motivation: Bias in Real-World Data

- Vaccine effectiveness
- SARS-CoV-2 infection and Long COVID
- Cancer therapies
- …

- Residual Bias in Observational Research
  - Unmeasured confounding
  - Measurement error
  - Selection bias
  - Missing data
  - …



Negative Controls: Current Frameworks

- Negative control outcome (NCO)
  - A clinical outcome that should not be causally affected by the treatment of interest
  - share similar sources of bias as the primary outcome
- Bias detection
- Bias correction

Research Article

Statistics in Medicine

Received 12 November 2012, Accepted 3 July 2013 Published online 30 July 2013 in Wiley Online Library

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

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

Martijn J. Schuemie,[a,b*†] Patrick B. Ryan,[b,c] William DuMouchel,[b,d] Marc A. Suchard[b,e] and David Madigan[b,f]

PNAS

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

Martijn J. Schuemie[a,b,1], George Hripcsak[a,c,d], Patrick B. Ryan[a,b,c], David Madigan[a,e], and Marc A. Suchard[a,f,g,h]

[a]Observational Health Data Sciences and Informatics, New York, NY 10032; [b]Epidemiology Analytics, Janssen Research & Development, Titusville, NJ 08560; [c]Department of Biomedical Informatics, Columbia University, New York, NY 10032; [d]Medical Informatics Services, New York–Presbyterian Hospital, New York, NY 10032; [e]Department of Statistics, Columbia University, New York, NY 10027; [f]Department of Biomathematics, University of California, Los Angeles, CA 90095; [g]Department of Biostatistics, University of California, Los Angeles, CA 90095; and [h]Department of Human Genetics, University of California, Los Angeles, CA 90095

Example of Implementation

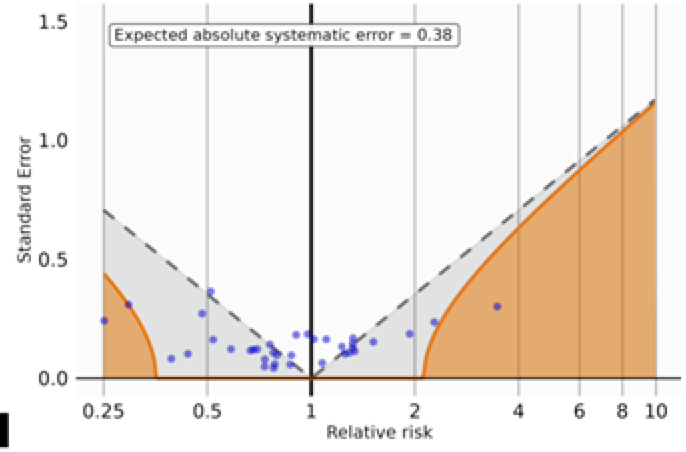
- Exposure: COVID-19 vaccination
- Outcome: SARS-CoV-2 infection

Annals of Internal Medicine

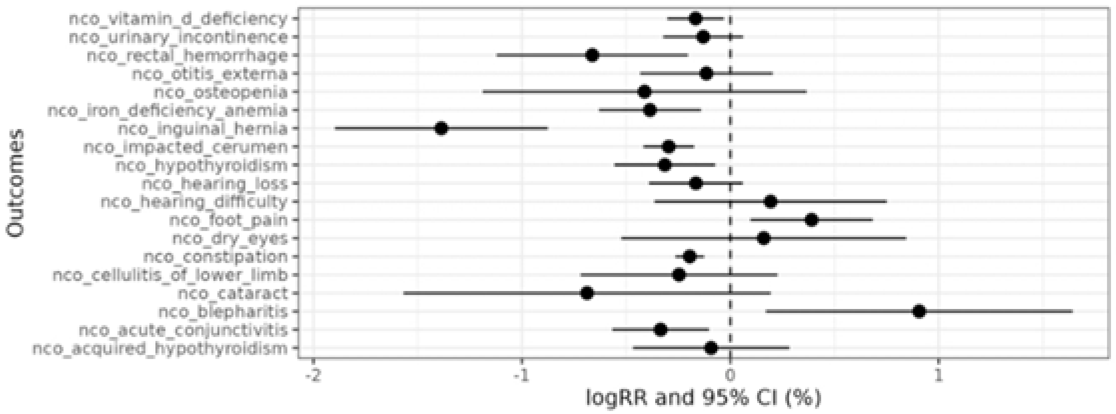
ORIGINAL RESEARCH

Real-World Effectiveness of BNT162b2 Against Infection and Severe Diseases in Children and Adolescents

Qiong Wu, PhD*; Jiayi Tong, MS*; Bingyu Zhang, MS; Dazheng Zhang, MS; Jiajie Chen, PhD; Yuqing Lei, MS; Yiwen Lu, BS; Yudong Wang, PhD; Lu Li, BA; Yishan Shen, MS; Jie Xu, PhD; L. Charles Bailey, MD, PhD; Jiang Bian, PhD; Dimitri A. Christakis, MD, MPH; Megan L. Fitzgerald, PhD; Kathryn Hirabayashi, MPH; Ravi Jhaveri, MD; Alka Khaitan, MD; Tianchen Lyu, MS; Suchitra Rao, MBBS, MSCS; Hanieh Razzaghi, PhD, MPH; Hayden T. Schwenk, MD, MPH; Fei Wang, PhD; Margot I. Gage Witvliet, PhD; Eric J. Tchetgen Tchetgen, PhD; Jeffrey S. Morris, PhD†; Christopher B. Forrest, MD, PhD†; and Yong Chen, PhD†



Categories	Examples
Infectious and parasitic diseases	Impetigo, Tinea capitis, Tinea corporis, Insect bite
Diseases of the skin and subcutaneous tissue	Contact dermatitis, Diaper rash, Acne
Diseases of the musculoskeletal system and connective tissue	Dislocations (Displacements - bone), Closed fracture of distal end of radius, Sprain of ankle, Scoliosis, Foot pain, Injury of free lower limb, Injury of upper extremity, Injury of right leg, Injury o left leg, Injury of right foot
Diseases of the nervous system	Seizure, Epilepsy, Concussion, Closed injury of head
Diseases of the eye and adnexa	Visual testing abnormal, Myopia, Astigmatism
Diseases of the ear and mastoid process	Wax in ear/impacted cerumen, Foreign body in ear
Diseases of the respiratory system	Snoring/Obstructive sleep apnea
Diseases of the digestive system	Umbilical hernia, Inguinal hernia
Endocrine, nutritional, and metabolic diseases	Obesity
Diseases of the speech and voice	Speech delay, Speech dysfunction, Tongue tie

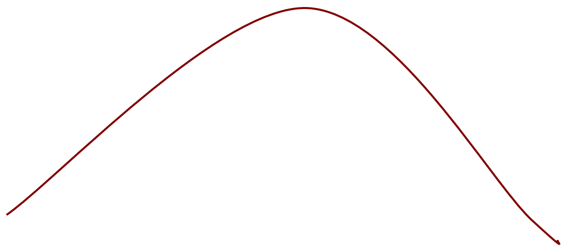


But… NCOs May Be Invalid

- Current frameworks assume all NCOs are valid
  - Normal-normal (N-N) model

$$y_i \sim N(\theta_i, s_i^2)$$

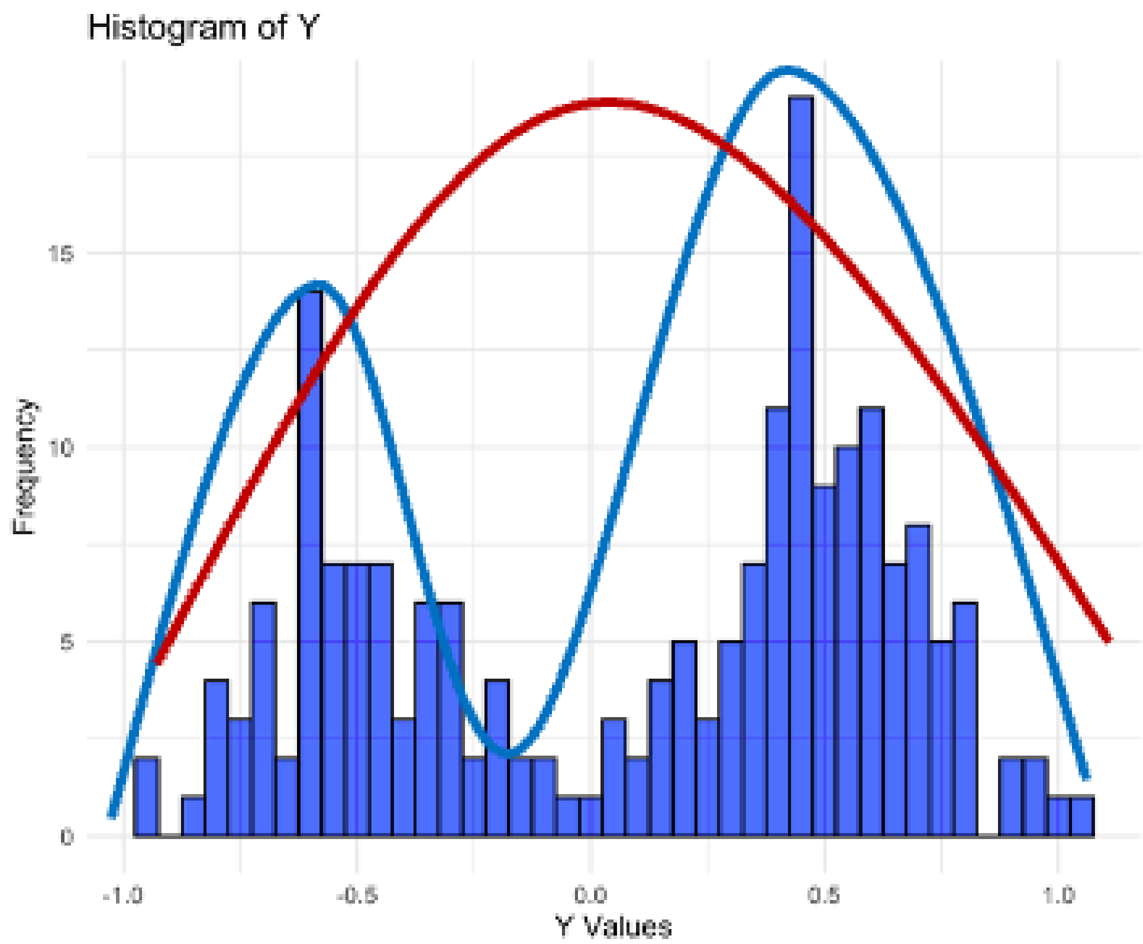
$$\theta_i \sim N(\mu, \sigma^2)$$



- In real-world scenarios, some NCOs may actually be invalid
  - Different confounding structures, data quality issues, coding practices, …
- This can bias bias-correction!

Why a Single Normal Fails

- Biased mean
- Larger variance

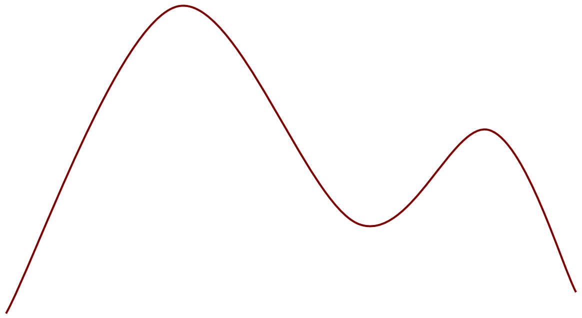


Proposed Method: Robustify P-value Calibration

- (A1) Two cluster mixture model
  - Relax the normality assumption
  - Mixture normal-normal (MN-N) model
- (A2) Majority rule
  - >50% NCOs are valid

$$y_i \sim N(\theta_i, s_i^2)$$

$$\theta_i \sim N(\mu, \sigma^2)$$



$$y_i \sim N(\theta_i, s_i^2)$$

$$\theta_i \sim \pi \cdot N(\mu_1, \sigma_1^2) + (1-\pi) \cdot N(\mu_2, \sigma_2^2)$$

$$\pi > 0.5$$

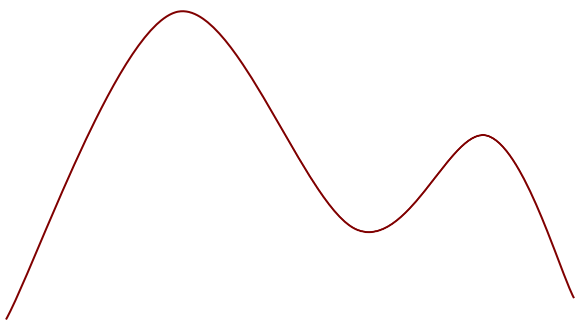
Mixture Model Framework

- For each NCO, observe estimated treatment effect y_i with standard error s_i
- Assume each NCO comes from one of the following two distributions:
  - Valid NCOs (true nulls)

$$y_i \sim N(\mu_1, \sigma_1^2 + s_i^2)$$

  - Invalid NCOs

$$y_i \sim N(\mu_2, \sigma_2^2 + s_i^2)$$



- Model the observed NCO distribution as a mixture:

$$\begin{cases} f(y_i) = \pi \cdot N(y_i|\mu_1, \sigma_1^2 + s_i^2) + (1-\pi) \cdot N(y_i|\mu_2, \sigma_2^2 + s_i^2) \\ \pi > 0.5 \end{cases}$$

- Estimate parameters $(\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ using EM algorithm

Calibrated p-value

- Using the estimated valid null distribution, for an effect estimate from a new drug-outcome pair, the two-sided p-value is then

Estimated mean of majority NCOs (valid)

$$p_{cal} = 2 \cdot \Phi\left(-\frac{|y_{n+1} - \boxed{\hat{\mu}_1}|}{\sqrt{\boxed{\hat{\sigma}_1^2} + s_{n+1}^2}}\right)$$

Estimated sd of majority NCOs (valid)

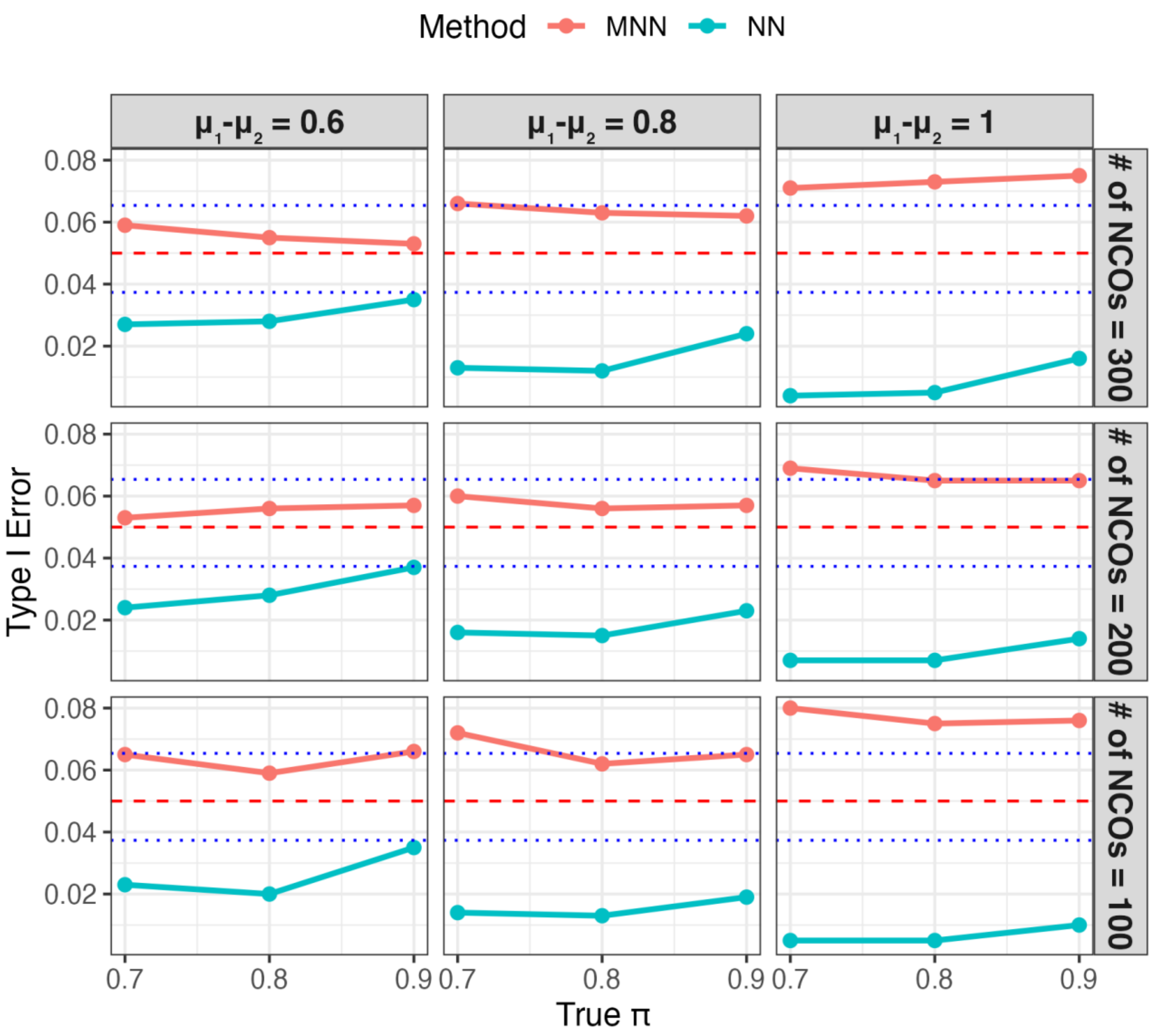
- Φ is the cumulative distribution function of the standard normal distribution

Simulation

- Proportion of valid NCO π: 0.7, 0.8, 0.9
- Number of NCOs n: 100, 200, 300
- Separation between valid and invalid means $\mu_1 - \mu_2$: 0.6, 0.8, 1.0

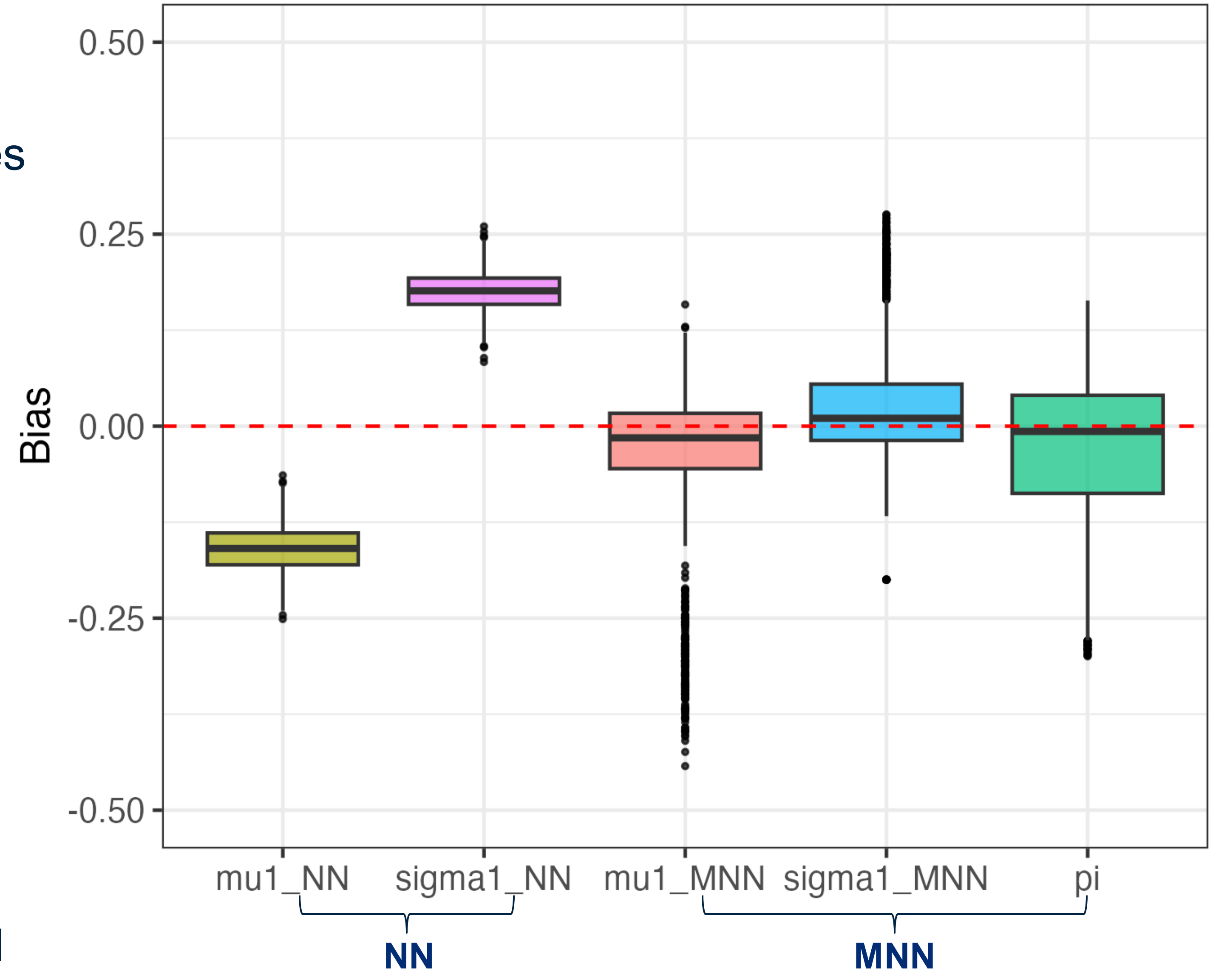
Simulation: Type I Error

- MNN achieved the nominal type I error



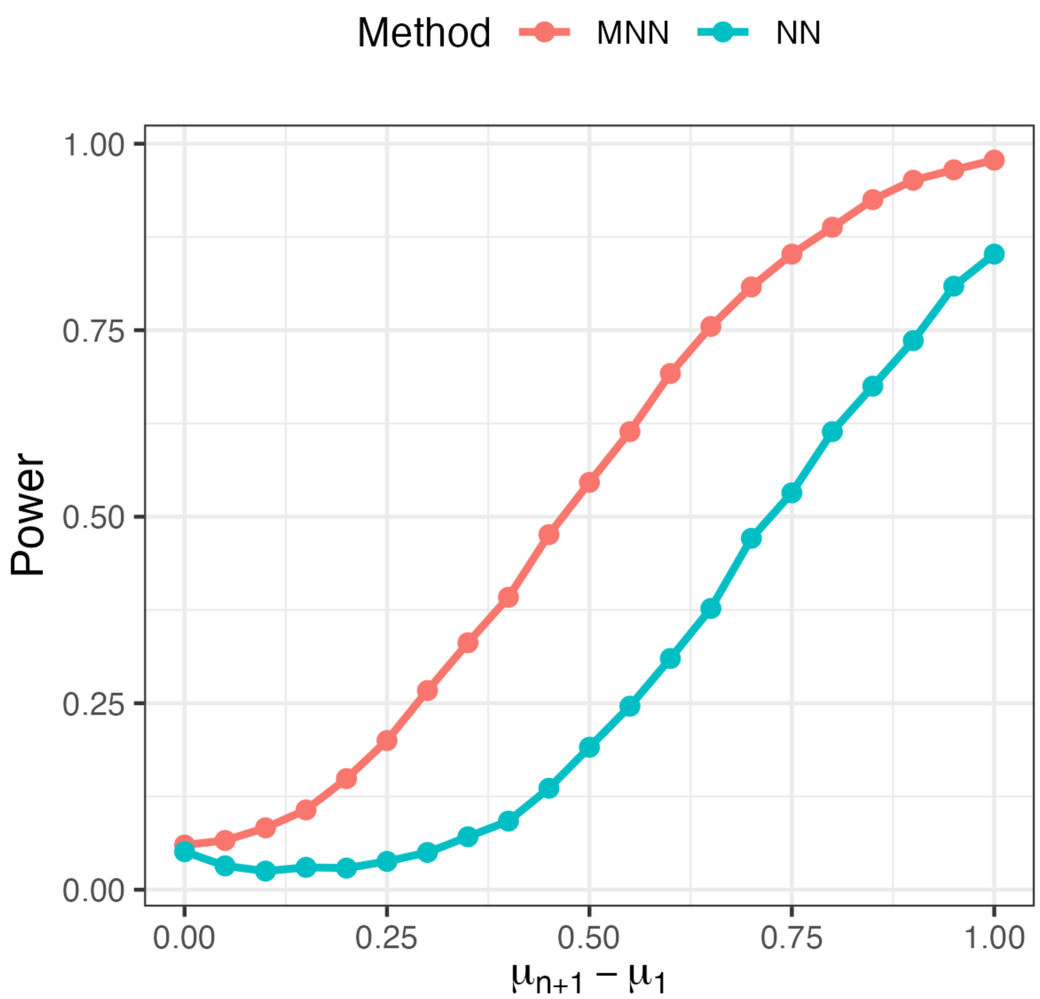
Simulation: Parameter Estimation

- MNN produced less biased estimates



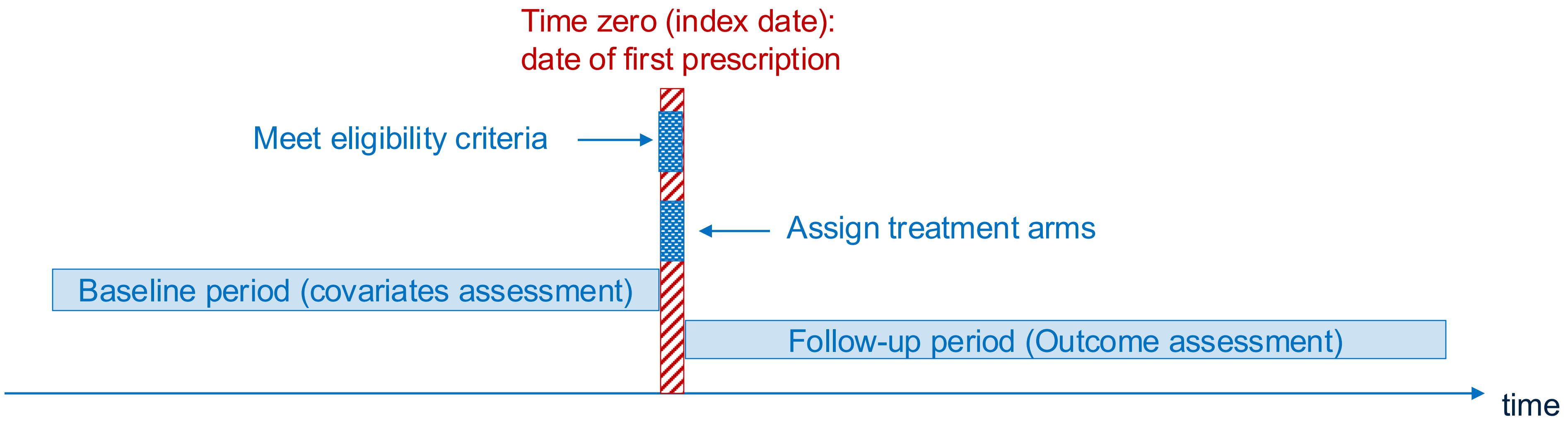
Simulation: Power

- MNN had higher power

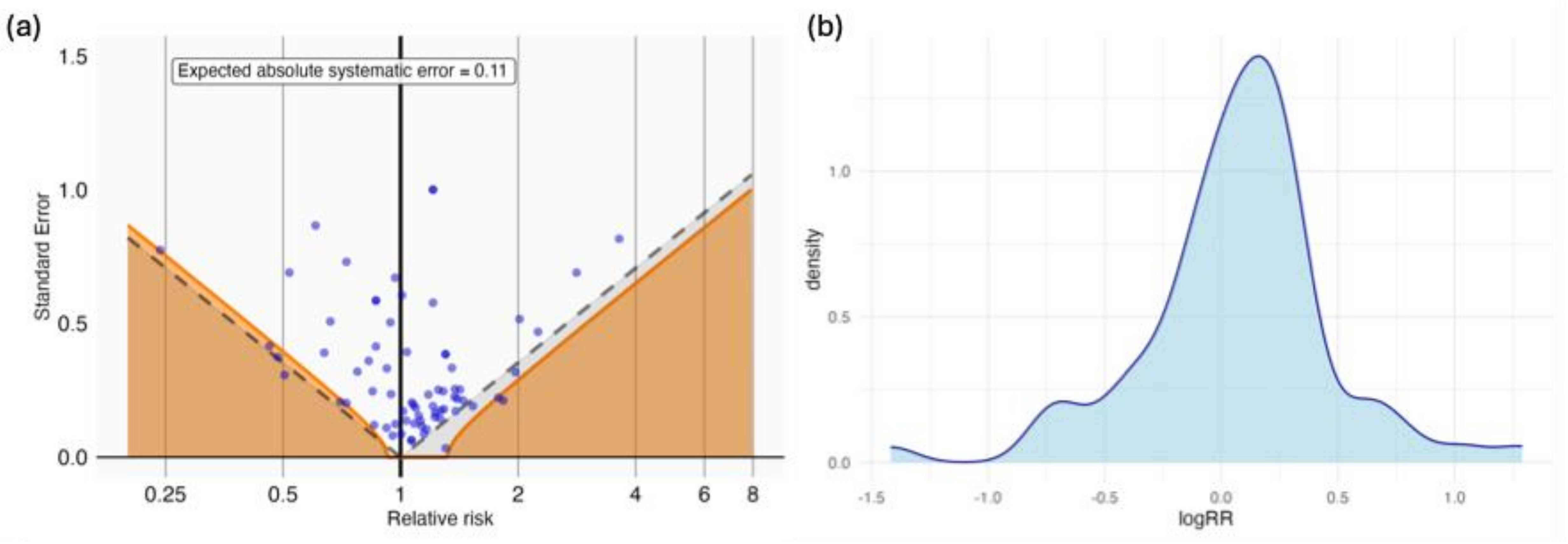


Real-World Use Case

- Data source: Penn Medicine EHR data
- Population: Patients with type 2 diabetes
- Treatment: GLP-1 receptor agonists
- Comparison: DPP4 inhibitors
- Outcomes: six cardiovascular outcomes
- Statistical analysis: large-scale propensity score matching + modified Poisson regression model

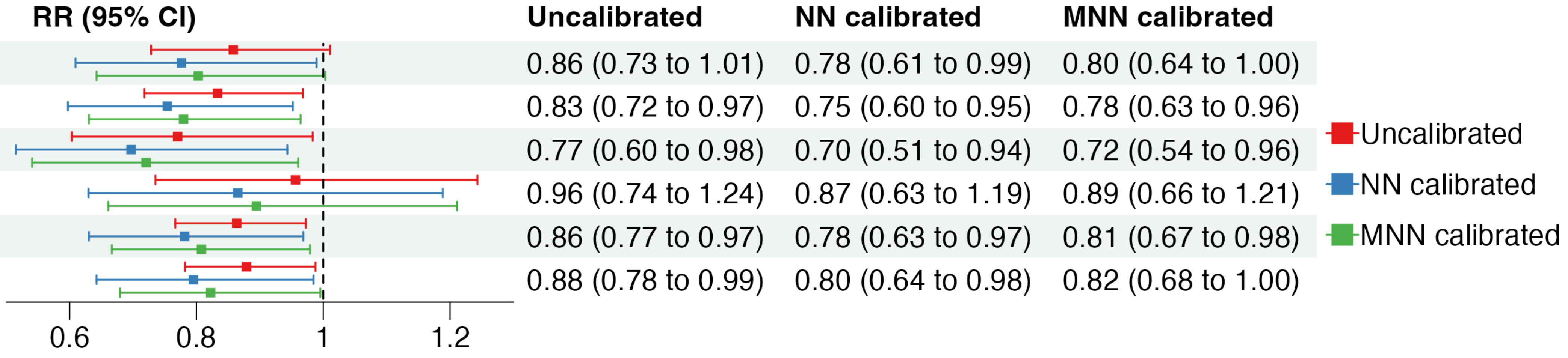


Distribution of NCOs



Treatment Effectiveness

Outcome Name	Uncalibrated	NN calibrated	MNN calibrated
Non-fatal MI	0.86 (0.73 to 1.01)	0.78 (0.61 to 0.99)	0.80 (0.64 to 1.00)
Non-fatal stroke	0.83 (0.72 to 0.97)	0.75 (0.60 to 0.95)	0.78 (0.63 to 0.96)
Hospitalization for UA	0.77 (0.60 to 0.98)	0.70 (0.51 to 0.94)	0.72 (0.54 to 0.96)
CV death	0.96 (0.74 to 1.24)	0.87 (0.63 to 1.19)	0.89 (0.66 to 1.21)
3-point MACE	0.86 (0.77 to 0.97)	0.78 (0.63 to 0.97)	0.81 (0.67 to 0.98)
4-point MACE	0.88 (0.78 to 0.99)	0.80 (0.64 to 0.98)	0.82 (0.68 to 1.00)



- MNN: smaller bias correction, narrower CI
- GLP1RAs have protective cardiovascular effects compared to DPP4is

Conclusion

- RWD enables large-scale observational research but is vulnerable to residual bias
- NCOs are essential tools but their validity cannot be guaranteed
- We propose a robust two-cluster model that:
  - Distinguishes valid from invalid NCOs
  - Enables bias correction even with partially invalid controls
  - Improves the reliability of p-values and confidence intervals

Acknowledgements

- Dazheng Zhang
- Huiyuan Wang
- Wenjie Hu
- Qiong Wu
- Chongliang Luo
- Lu Li
- Tsai Hor Chan
- Yudong Wang
- Yuru Zhu
- Martijn Schuemie
- Patrick Ryan
- George Hripcsak
- Marc Suchard
- Yong Chen

Contact:

Dr. Yong Chen: ychen123@pennmedicine.upenn.edu

Bingyu Zhang: bingyuz7@sas.upenn.edu