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

Bingyu Zhang[1,2], Dazheng Zhang, PhD[1,3], Huiyuan Wang, PhD [1,3], Wenjie Hu, PhD [1,3], Qiong Wu, PhD [1,3,4], Chongliang Luo, PhD[5], Lu Li[1,2], Tsai Hor Chan[1,3,6], Yudong Wang, PhD[1,3], Martijn Schuemie[7,8,9], Patrick Ryan[7,8,10], George Hripcsak, MD[7,10], Marc Suchard[7,9,11], Yong Chen, PhD[1,2,3]

[1] The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
[2] Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
[3] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4] Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA
[5] Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO, USA
[6] Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China
[7] Observational Health Data Sciences and Informatics, New York, NY, USA
[8] Global Epidemiology Organization, Johnson & Johnson, Titusville, NJ, USA
[9] Department of Biostatistics, University of California, Los Angeles, CA, USA
[10] Department of Biomedical Informatics, Columbia University Irving Medical Center, New York City, NY, USA
[11] VA Informatics and Computing Infrastructure, US Department of Veterans Affairs, Salt Lake City, UT, USA

Background

Observational studies play an increasingly pivotal role in evaluating the safety and effectiveness of medical interventions, especially in real-world settings where randomized controlled trials (RCTs) may be infeasible, costly, or ethically impractical. However, unlike RCTs, observational studies are inherently prone to systematic error, including unmeasured confounding, selection bias, and measurement error. These sources of bias can distort effect estimates and inflate the rate of false-positive findings, even when traditional p-values suggest statistical significance.

To mitigate this, empirical calibration has emerged as a powerful method that uses negative control outcomes (NCOs)—outcomes for which no causal association with the treatment is expected.[1,2] By modeling the null distribution empirically, researchers can better account for systematic error, improving the validity of statistical inference in observational research.

Nevertheless, a critical assumption in current empirical calibration methods is that all NCOs are valid, that is, their true effect sizes arise from a single, shared normal distribution. In practice, some NCOs may be affected by residual confounding, measurement error, or other biases that cause their estimated effects to deviate from the null. Ignoring such “invalid” NCOs can lead to biased null distribution estimates and miscalibrated p-values and poor type I error control, especially when the presence of invalid controls is non-negligible.

This paper introduces a robust extension that accommodates the presence of two clusters of NCOs: a dominant cluster of valid controls and a minority cluster of potentially invalid ones. By modeling the empirical null as a mixture distribution, our approach aims to achieve better calibration of p-values, ensuring more reliable inference even in the presence of heterogeneity among NCOs.

Methods

PROBLEM DEFINITION

Let y_i denote the estimated effect size (e.g., log relative risk) for the i-th NCO, with standard error s_i. We assume that NCOs arise from a two-component Gaussian mixture, representing either:

- Valid NCOs ("true nulls"):

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

- Invalid NCOs:

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

Let π denote the proportion of valid NCOs, assumed to be greater than 0.5.

MIXTURE MODEL FRAMEWORK

We model the observed distribution of NCOs as:

$$f(y_i) = \pi \cdot N\left(y_i|\mu_1, \sigma_1^2 + s_i^2\right) + (1-\pi) \cdot N\left(y_i|\mu_2, \sigma_2^2 + s_i^2\right).$$

Model parameters $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi)$ are estimated using the Expectation-Maximization (EM) algorithm. In the E-step, posterior probabilities of each NCO being valid are computed. In the M-step, parameters are updated to maximize the likelihood given these responsibilities.

CALIBRATED P-VALUES

After estimating the distribution of valid NCOs $N(\hat{\mu}_1, \hat{\sigma}_1^2)$, we compute calibrated two-sided p-values for a new effect estimate y_{n+1} with standard error s_{n+1} from a new drug-outcome pair:

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

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Results

SIMULATION STUDIES

We conducted simulation experiments to assess the performance of the proposed Mixture Normal-Normal (MNN) calibration method under varying conditions. Each scenario was replicated 1,000 times. Key evaluation metrics include:

1) Estimation accuracy of the model parameters μ_1, σ_1, π
2) Type I error control for calibrated p-values under the null
3) Statistical power under increasing signal strength

We varied the proportion of valid NCOs ($\pi \in \{0.7, 0.8, 0.9\}$), the separation between valid and invalid means ($\mu_1 - \mu_2 \in \{0.6, 0.8, 1.0\}$), and the number of NCOs (100, 200, 300). For each replication, standard errors were drawn from a plausible distribution based on real-world EHR-based studies.

MNN consistently achieved nominal Type I error across most scenarios, while the baseline Normal-Normal (NN) method often failed. Parameter estimates showed lower bias and variance under MNN. In power analysis, MNN demonstrated uniformly higher sensitivity in detecting non-null signals as the separation from the null increased.

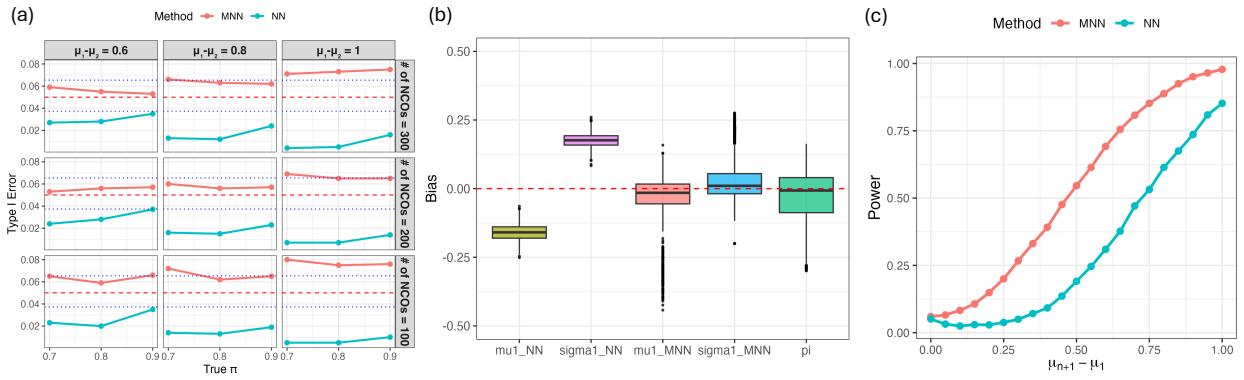

Figure 1. Simulation results comparing the proposed Mixture Normal-Normal (MNN) method and the baseline Normal-Normal (NN) method. (a) Type I error rates at $\alpha = 0.05$ across varying values of π, effect separations, and numbers of NCOs. The dashed red and blue lines denote the nominal 0.05 level. (b) Bias in parameter estimates of μ_1, σ_1, π when true $\pi = 0.8$ and $\mu_1 - \mu_2 = 0.8$. (c) Power curves showing the probability of detecting non-null signals as the effect difference $(\mu_{n+1} - \mu_1)$ increases.

REAL-WORLD USE CASE

To evaluate the practical utility of our proposed calibration framework, we applied it to a real-world comparative effectiveness study using electronic health record (EHR) data from Penn Medicine. The analysis focused on the cardiovascular effects of glucagon-like peptide-1 receptor agonists (GLP-1RA) versus dipeptidyl peptidase-4 inhibitors (DPP4i)—two commonly prescribed classes of antidiabetic medications.

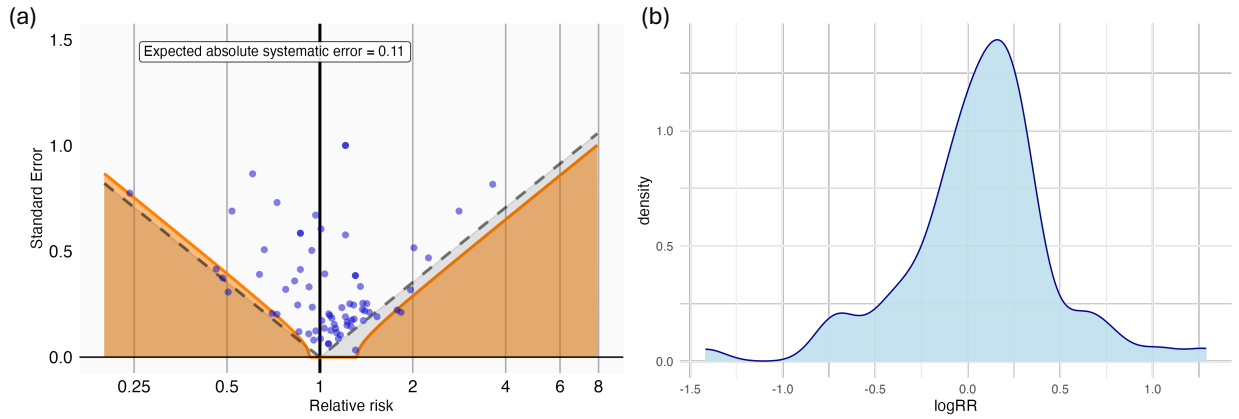

Figure 2. Negative control outcomes distribution. (a) Scatter plots of relative risk and standard error estimates. (b) Density of the relative risk estimate distribution.

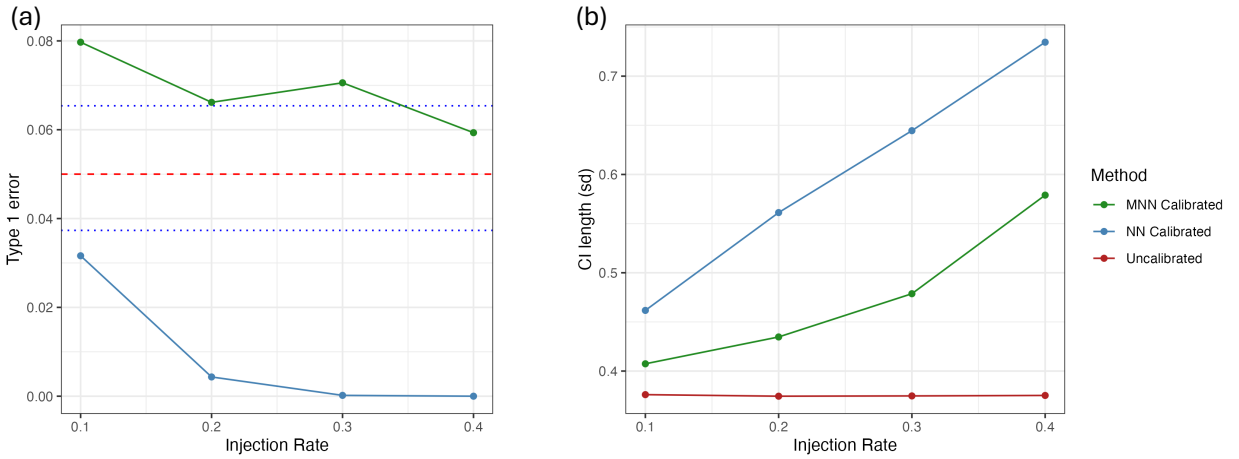

Figure 3. Evaluation of calibration methods. Results over 100 replications across varying proportions of injected invalid NCOs (10% to 40%). (a) Type I error rates for NN and MNN. For each injection rate, a proportion of NCOs are modified, while the remaining NCOs are used in the leave-one-out cross-validation (LOOCV) framework to assess type I error. (b) Average confidence interval length for uncalibrated, NN calibrated, and MNN calibrated estimates.

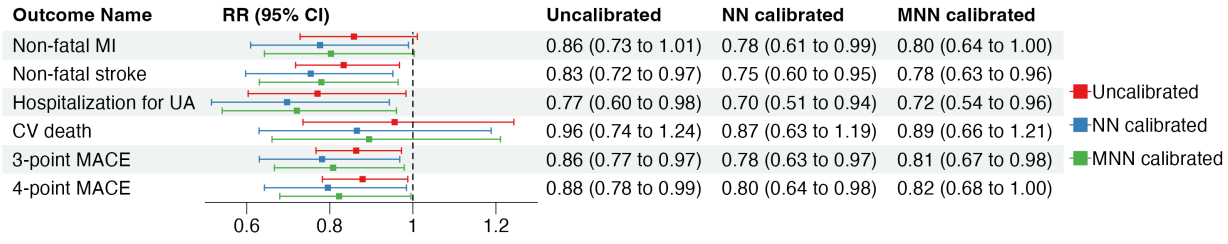

Figure 4. Estimated relative risks for cardiovascular outcomes comparing GLP-1RA vs DPP4i. Comparative effectiveness estimates across six cardiovascular endpoints. Error bars represent 95% confidence intervals under three methods: uncalibrated, NN calibrated, and MNN calibrated. The MNN approach yields both improved error control and tighter confidence bounds in most outcomes.

Conclusion

This study advances the empirical calibration framework by introducing a flexible mixture model that explicitly accounts for the presence of partially invalid NCOs, a common but under-addressed challenge in real-world observational research. By modeling NCOs as arising from a two-component Gaussian mixture, our method offers improved robustness in bias estimation and p-value calibration.

Through extensive simulation studies and application to electronic health record data, we demonstrate that the proposed MNN method outperforms existing approaches in type I error control, parameter estimation accuracy, and statistical power. These findings underscore the importance of relaxing the overly strict assumption that all NCOs are valid.

From a practical standpoint, several considerations emerge. First, the estimated proportion of valid NCOs (π) provides an interpretable diagnostic: consistently low estimates should be viewed as a signal to revisit the choice of NCOs, while intermediate values (e.g., close to 0.5) may warrant sensitivity analyses to assess robustness. Second, when extending to multi-site or federated studies, the framework can be integrated with meta-analytic or hierarchical calibration strategies, thereby accommodating variation in NCO quality across heterogeneous databases. Together, these recommendations provide a roadmap for applying the MNN method in practice, ensuring more reliable inference in both single-site and distributed observational studies.

References

1. Schuemie MJ, Ryan PB, DuMouchel W, et al. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med* 2014; 33: 209–218.
2. Schuemie MJ, Hripcsak G, Ryan PB, et al. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences* 2018; 115: 2571–2577.