Calibrated Bayesian NI Inference for Paired Binary Endpoints

- Halim Halidola, BioNTech SE

- 21-Oct-2026

BIONTECH

Motivation

- Non-inferiority (NI) studies with paired binary endpoint: devices, diagnostics, crossover studies…
- Inference driven by discordant pairs.
- Sparse/zero discordance causes instability for asymptotic methods.
- Need robust approach with Type I error control.

CUHF study example:
The primary objective is to demonstrate the test product is NI to the reference product in terms of success rate.
Difference (Reference - Test) in success rate: $P_{ref} - P_{test} = (P_{11} + P_{01}) - (P_{11}+P_{10}) = P_{01} - P_{10}$

$$H_0: P_{01} - P_{10} \geq M$$
$$H_1: P_{01} - P_{10} < M$$

where M is the NI margin.

Table 1. 2×2 Contingency Table for Paired Binary Outcomes

	Reference Success	Reference Failure	Total
Test Success	a (p_{11})	b (p_{10})	*a+b* (P_{test})
Test Failure	c (p_{01})	d (p_{00})	*d+d*
Total	*a+c* (P_{ref})	*d+d*	*n*

Correlation coefficient

$$\rho = \frac{p_{11} - p_{test}p_{ref}}{\sqrt{p_{test}(1-p_{test})p_{ref}(1-p_{ref})}}$$

Method – Calibrated Bayesian NI Framework

Bayesian model

$$(p_{11}, p_{10}, p_{01}, p_{00}) \sim Dirichlet(0.5,\ 0.5,\ 0.5,\ 0.5)$$

$$(a, b, c, d) \mid p \sim Multinomial(n, p)$$

By conjugacy,

$$p \mid (a, b, c, d) \sim Dirichlet(a + 0.5,\ b + 0.5,\ c + 0.5,\ d + 0.5).$$

The non-inferiority parameter is

$$\Delta = p_{01} - p_{10}.$$

Bayesian decision rule

Declare non-inferiority if

$$Pr(\Delta < M \mid Data) > \gamma$$

Calibration

Choose γ^* such that

$$P(\text{Declare NI} \mid \Delta = M) \leq \text{target alpha},$$

thereby controlling the one-sided frequentist Type I error rate at nominal level.

Evaluate OC: Type I error rate, Power, Coverage probability

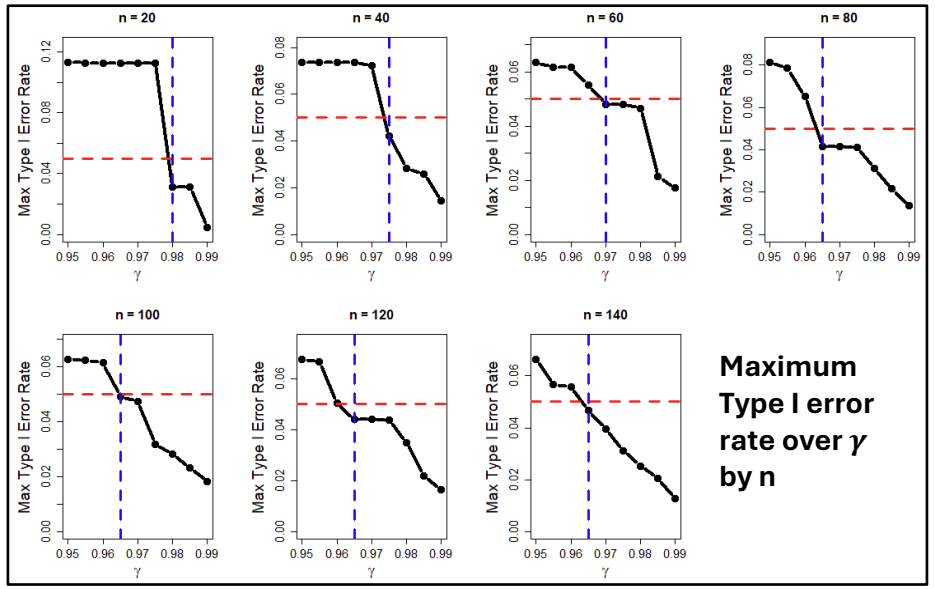
Frequentist Benchmark: Tango Method

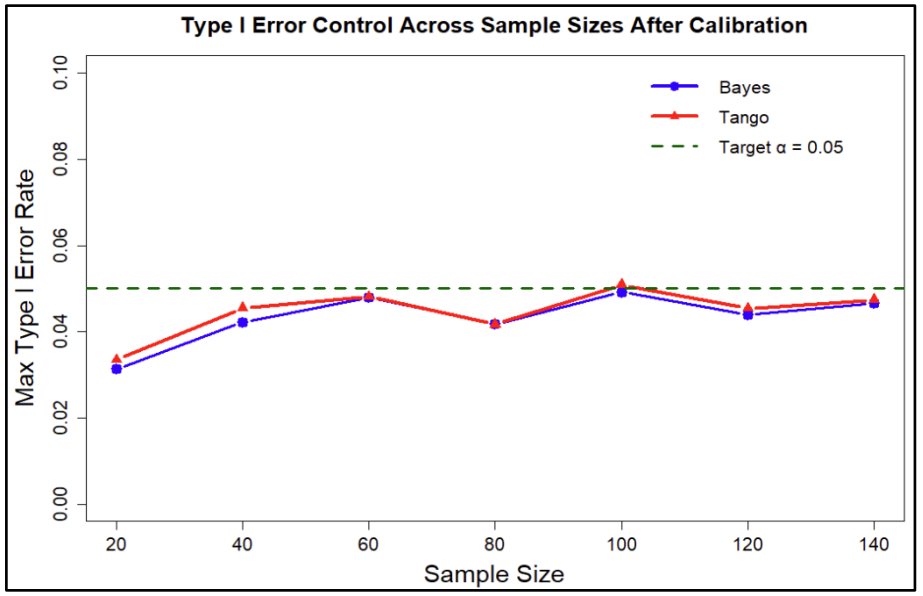
- Treat $\Delta = p_{01} - p_{10}$ as the parameter of interest.
- For each candidate value of Δ, estimates the total discordance $\psi = p_{10} + p_{01}$.
- Construct a score test using that constrained model, and obtain the confidence interval by collecting all values of Δ that are not rejected, i.e.,
- $Z(\Delta)^2 \leq \chi^2_{1,\,1-\alpha}$
- Tango method generally provides more reliable inference than Wald-type methods when discordant pair counts are sparse or zero.

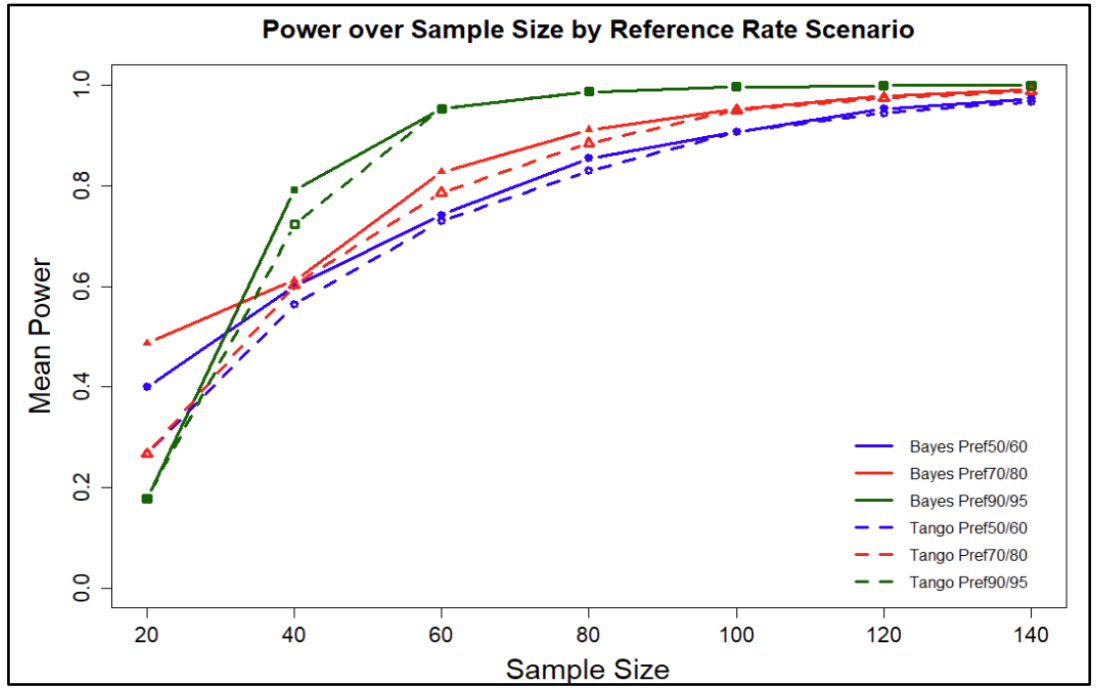
Simulation Results

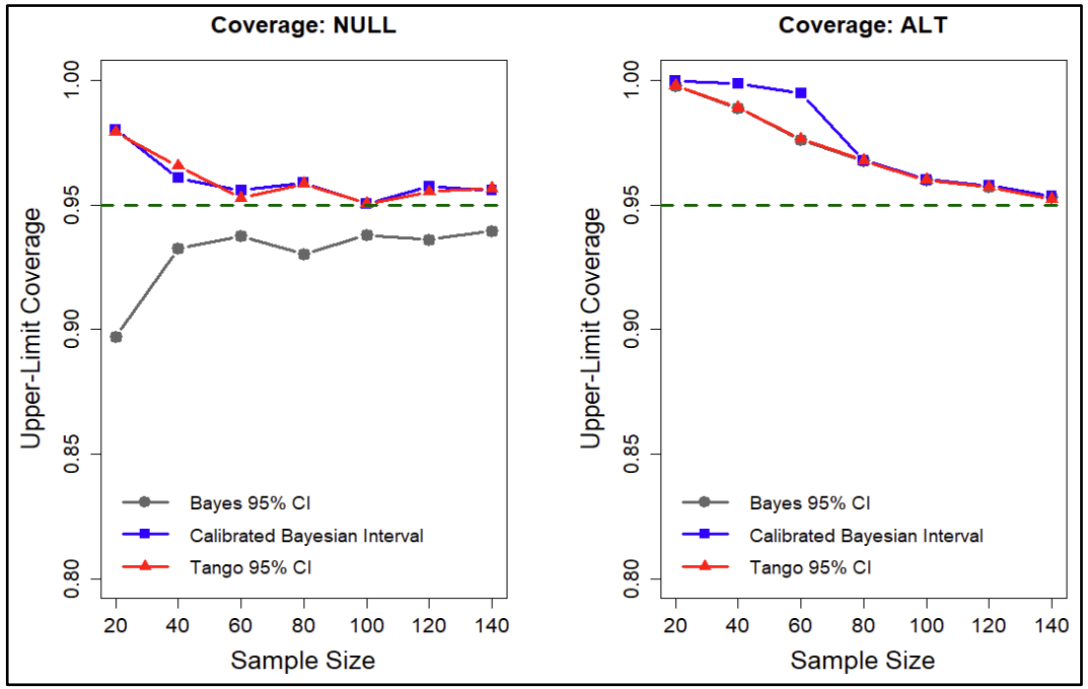
-- Calibration and Type 1 Error Rate

- The conventional threshold γ = 0.95 did not consistently achieve nominal Type I error control.
- Calibration identified scenario-specific thresholds γ* that maintained a one-sided Type I error rate near 5%.
- Across all investigated reference rate and correlation coefficient scenarios, γ* was remarkably stable (approximately 0.96-0.98).
- Variability in γ* was driven primarily by sample size and discreteness rather than underlying marginal rates.
- After calibration, Bayesian and Tango methods achieved very similar Type I error performance.









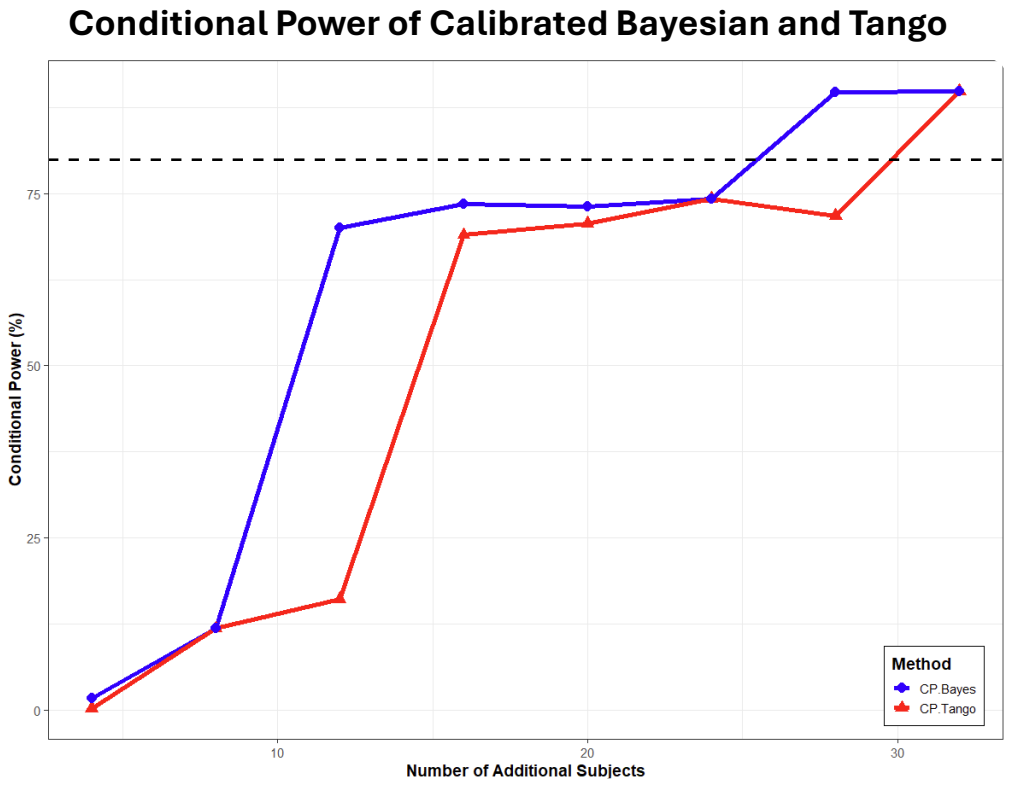
Simulation Results

-- Power and Coverage

- After calibration, the Bayesian procedure achieved power comparable to Tango, with a slight power advantage at smaller sample sizes.
- Bayes calibrated intervals showed similar coverage as Tango intervals.
- Under the null hypothesis, the uncalibrated Bayesian credible interval showed persistent undercoverage, highlighting the benefit of calibration.

Simulation Results

-- Conditional Power Assessment



CUHF study example

- Simulation-based reassessment using pooled nuisance parameter estimates from the ongoing study.
- Conditional power was evaluated for Tango and calibrated Bayesian approaches across additional sample sizes.
- Calibrated Bayesian analysis yielded comparable, and occasionally higher, conditional power than Tango.

Note: Conditional power increased in a stepwise manner due to the **discrete nature of the paired binary endpoint**, where small sample size changes can shift many simulated studies from failing to meeting the non-inferiority criterion. This reflects a property of the endpoint and analysis method rather than a meaningful change in performance.

Conclusions

- A calibrated Bayesian framework was developed for NI inference with paired binary endpoints and sparse discordant pairs.
- Calibration of the Bayesian decision threshold achieved one-sided Type I error control near the nominal level and produced operating characteristics comparable to the Tango method across the investigated scenarios
- Small power gains were observed for the calibrated Bayesian approach in some small-sample settings.
- Calibration substantially improved interval performance relative to the uncalibrated Bayesian approach.
- The proposed method provides a Bayesian alternative with controlled frequentist operating characteristics for paired binary NI studies.

References

- Tango, T. (1998), Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Statist. Med., 17: 891-908. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B
- Yang P., Wang L., and Yuan Y. (2026), On the Calibration of Bayesian Success Criteria and Operating Characteristics for Clinical Trials. arXiv preprint arXiv:2603.20015. https://doi.org/10.48550/arXiv.2603.20015
- US Food and Drug Administration. (2026) Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products: guidance for Industry: draft guidance. Rockville, MD: US Food and Drug Administration.
- US Food and Drug Administration. (2017) Comparative Analyses and Related Comparative Use Human Factors Studies for a Drug-Device Combination Product Submitted in an ANDA: Draft Guidance for Industry. Rockville, MD: US Food and Drug Administration.
- Yang, H., & Novick, S. (2019). Bayesian Analysis with R for Drug Development: Concepts, Algorithms, and Case Studies (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315100388

Questions or Comments?

