# A Unified Inference Framework for RD and RR in Small-Sample and Low-Incidence Binary Endpoints in Clinical Trials

**Linbo Wang**

University of Toronto


Supported by BARDS ARC funding from Merck Sharp & Dohme LLC

PSI 2026

June 16, 2026

# Acknowledgements


## Collaborators

Jingxin Yan, X. Gregory Chen, Margarita Donica, Yujie Zhao, Larry Leon, Qizhai Li, and Thomas Richardson.

## Funding

This work was supported by BARDS Academic Research Collaboration (ARC) funding from Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA:

*Improving Clinical and HTA Analysis for Binary Endpoints: Rethinking about RD and RR Methodologies*.

MSD


*The views expressed are my own and do not represent the views of the University of Toronto or Merck Sharp & Dohme LLC.*

# Binary Endpoints in Clinical Trials

Small samples and rare events can make standard RD/RR inference unreliable for binary endpoints.

- Small samples: early-phase trials, subgroup analyses, pediatric or rare-disease studies.
- Rare events: oncology safety endpoints or low-incidence efficacy endpoints in subgroup analyses.
- Zero cells can appear even when the trial itself is not small.

# Binary Endpoints: RD and RR Both Matter

## Risk difference (RD)

- Absolute effect.
- Extra/fewer events per 100 patients.
- Descriptive safety summaries.

## Risk ratio (RR)

- Relative effect.
- Risk multiplier.
- Percent change in risk.

RD and RR answer complementary stakeholder questions; the method should be reliable on both scales.

Evidence synthesis and regulatory discussions may use either scale; conclusions should not depend on an arbitrary effect-measure choice.

# Findings and Main Target

**Main empirical finding**

Regular settings: similar.
Hard settings: methods separate and can fail.

- In common-event, moderate-sample settings, representative methods are close.
- In small-sample, low-incidence, or zero-event settings, standard methods can be unstable or miscalibrated.

Main target: reliable finite-sample inference for both RD and RR.

# Review: Common Methods Are Scale-Specific

| | RR scale | RD scale |
|---|---|---|
| Score / exact | CMH / exact intervals | MN / score intervals |
| Regression / GLM | Log-binomial / log-Poisson | LPM / robust SE |
| Bias reduction | Firth / bias-reduced GLM | Bias-reduced GLM |
| Sparse correction | Continuity / Beta | Continuity / Beta |

# Review: Common Methods Are Scale-Specific

| | RR scale | RD scale |
|---|---|---|
| Score / exact | CMH / exact intervals | MN / score intervals |
| Regression / GLM | Log-binomial / log-Poisson | LPM / robust SE |
| Bias reduction | Firth / bias-reduced GLM | Bias-reduced GLM |
| Sparse correction | Continuity / Beta | Continuity / Beta |

! **Fragmented Menu:** different RD/RR defaults affect consistency.

# Review: Common Methods Are Scale-Specific

| | RR scale | RD scale |
|---|---|---|
| Score / exact | CMH / exact intervals | MN / score intervals |
| Regression / GLM | Log-binomial / log-Poisson | LPM / robust SE |
| Bias reduction | Firth / bias-reduced GLM | Bias-reduced GLM |
| Sparse correction | Continuity / Beta | Continuity / Beta |

! **Fragmented Menu:** different RD/RR defaults affect consistency.

! **Poor Hard-Case Performance:** sparse data break methods.

# Proposed Solution: A Unified RD/RR Framework

Let

$$p_a(\boldsymbol{v}) = E(Y \mid A = a, \boldsymbol{V} = \boldsymbol{v}), \qquad a = 0, 1.$$

$$\mathrm{RR}(\boldsymbol{v}) = \frac{p_1(\boldsymbol{v})}{p_0(\boldsymbol{v})}, \qquad \mathrm{RD}(\boldsymbol{v}) = p_1(\boldsymbol{v}) - p_0(\boldsymbol{v}).$$

Following Richardson, Robins, and Wang (2017), model a target effect together with the odds-product nuisance parameter:

$$\theta(\boldsymbol{v}) = \log\{\mathrm{RR}(\boldsymbol{v})\} \quad \text{or} \quad \operatorname{arctanh}\{\mathrm{RD}(\boldsymbol{v})\}, \qquad \phi(\boldsymbol{v}) = \log\{\mathrm{OP}(\boldsymbol{v})\}.$$

One framework for both reporting scales.

# What We Add for Challenging Settings

## Bias reduction

Firth-type procedures improve finite-sample estimation stability.

## Exact intervals

Blaker-style exact confidence intervals improve calibration when asymptotics are weak.

## Zero-event correction

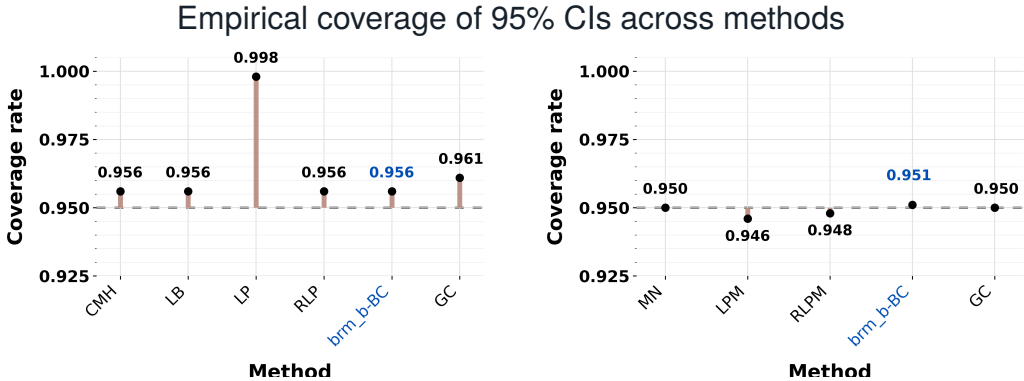
Bayesian correction generalizes continuity correction through Beta priors on event risks.

Recommended RD/RR analysis:
unified framework + finite-sample corrections.

# Simulation Design

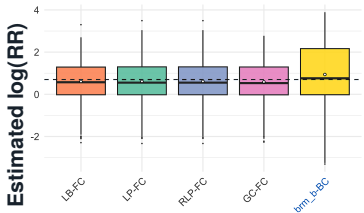
| Dimension | Levels | Purpose |
|---|---|---|
| Data-generating model | $V \sim U(0, 0.6)$; constant RD or RR | covariate-adjusted risks vary across patients |
| Incidence | common, rare | regular vs low-event settings |
| Sample size | $n = 50, 200, 500$ | small to moderate trials |
| Hypothesis | null, alternative | type I error and power |
| Effect scale | RD, RR | primary performance plus secondary agreement diagnostic |
| Metrics | estimate, SE, coverage, SE accuracy ratio | estimation, interval calibration, and SE calibration |

# Regular Settings: Representative Methods Look Similar

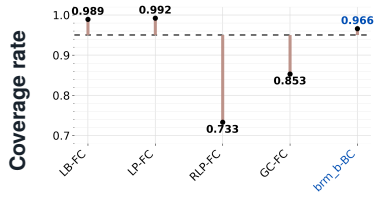



In common-event, moderate-sample settings, representative methods are close enough that method choice is rarely the main driver (except for Log-Poisson).

# Rare RR: Estimation and Coverage Separate the Methods



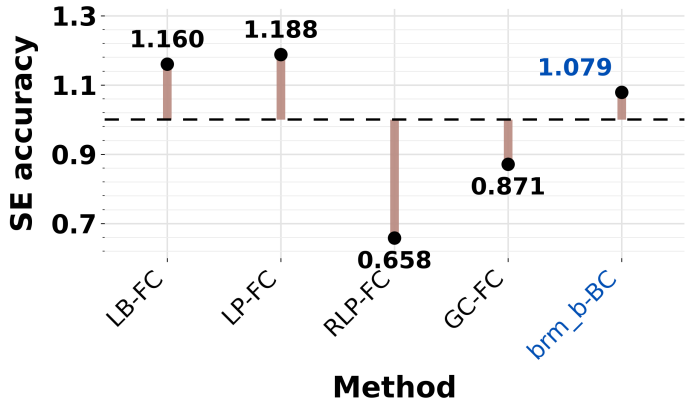

Estimate




Coverage

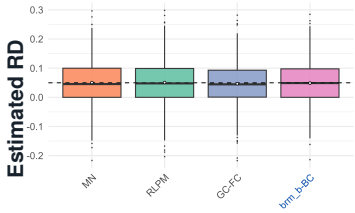
Sparse RR analysis separates methods sharply: stable point estimation is not enough if interval coverage is miscalibrated.

# Rare RR: SE Accuracy Checks Calibration

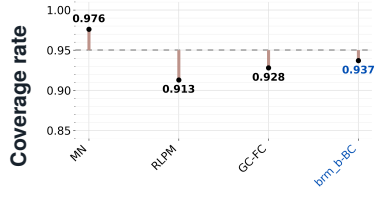



SE accuracy = estimated SE / Monte Carlo SE; values closer to 1 indicate better SE calibration.

# Rare RD: Reliability Issue on the Absolute Scale



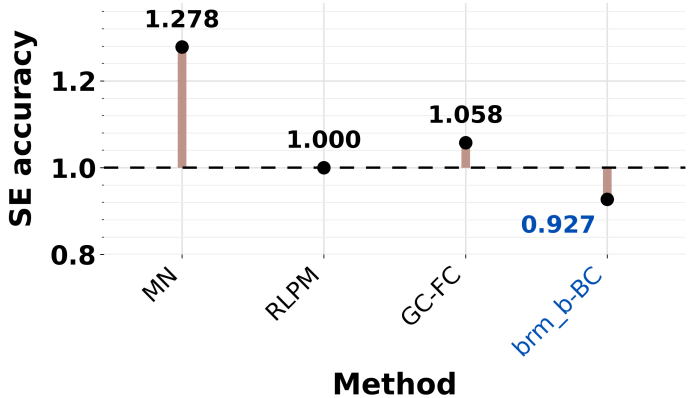

Estimate




Coverage

The same reliability issue appears on the absolute scale; the same default workflow applies to RD and RR.

# Rare RD: SE Accuracy Checks Scale Robustness




The same estimated SE / Monte Carlo SE diagnostic is used on RD, so good behavior is not specific to one reporting scale.

# Secondary Diagnostic: RD/RR Testing Agreement

**Setting:** $n = 200$, $E\{p_0(V)\} \approx 0.01$, $E\{p_1(V)\} \approx 0.11$ ($\mathrm{RD} = 0.10$, $\mathrm{RR} = 11$).

| Comparison | RR reject | RD reject | Both | RR only | RD only | Neither | Correct agree |
|---|---|---|---|---|---|---|---|
| CMH vs MN | 0.888 | 0.859 | 0.859 | 0.029 | 0.000 | 0.112 | 0.885 |
| LB vs LPM | 0.466 | 0.843 | 0.440 | 0.026 | 0.403 | 0.131 | 0.771 |
| RLP vs RLPM | 0.829 | 0.899 | 0.828 | 0.001 | 0.071 | 0.100 | 0.892 |
| brm RR vs brm RD | 0.483 | 0.903 | 0.481 | 0.002 | 0.422 | 0.095 | 0.835 |
| `brm_b-BC` RR vs RD | **0.908** | 0.894 | **0.887** | 0.021 | 0.007 | 0.085 | **0.913** |
| GC-FC RR vs RD | 0.874 | 0.892 | 0.869 | 0.005 | 0.023 | 0.103 | 0.894 |

Even when the alternative is far from the null on both scales, some method pairs still give scale-dependent conclusions.

# Take-Home Message

Default workflow: use `brm_b-BC` as the primary RD/RR analysis across settings.

## Implementation

The `brm_plus` R package is available on GitHub and implements the proposed methods for applied use.

`https://github.com/hta-pharma/brm_plus`

# References I


- Richardson, T. S., Robins, J. M., and Wang, L. (2017). On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 112(519), 1121–1130.
- Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*, 28(4), 783–798.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38.