17 June 2026

Beyond Dichotomization: Efficient Estimation of Response Rates using Continuous Outcomes

Michael Sweeting, GSK

PSI 2026

Acknowledgements

GSK Biostatistics:

Tom Drury

Michael Seath

Lindsey Schader

Jessica Lim

Stephen Weng

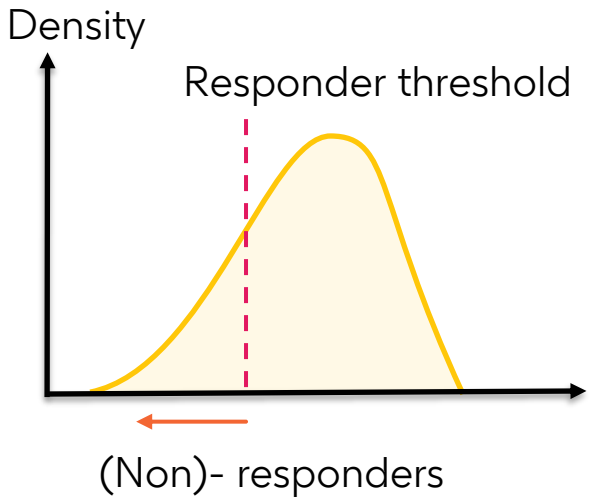
David Whitney

Paul Newcombe

Responder endpoints are common across medical research

- Some examples:

Disease Area	Diagnostic / responder threshold
Diabetes	HbA1c < 7.0%
Respiratory	Patient Reported Outcomes, e.g. CAT, SGRQ, E-RS, ACQ6
Oncology	RECIST: ≥ 30% reduction in tumour size Myelofibrosis: ≥ 35% reduction in spleen volume
Rheumatology	ACR20 ≥ 20%
Hypertension	SBP ≥ 140mmHg, DBP ≥ 90mmHg



- Review of 21,435 RCTs found that 66% have a binary primary efficacy outcome[1]

1.. van Zwet EW, Harrell Jr FE, Senn SJ. An Empirical Assessment of the Cost of Dichotomization of the Outcome of Clinical Trials. *Statistics in Medicine*. 2026;45(3-5):e70402. doi:10.1002/sim.70402

Change the estimator, not the estimand

- An alternative to dichotomization is to focus on estimands for continuous data
  - e.g. Comparison of means of the underlying continuous outcome (ANCOVA)

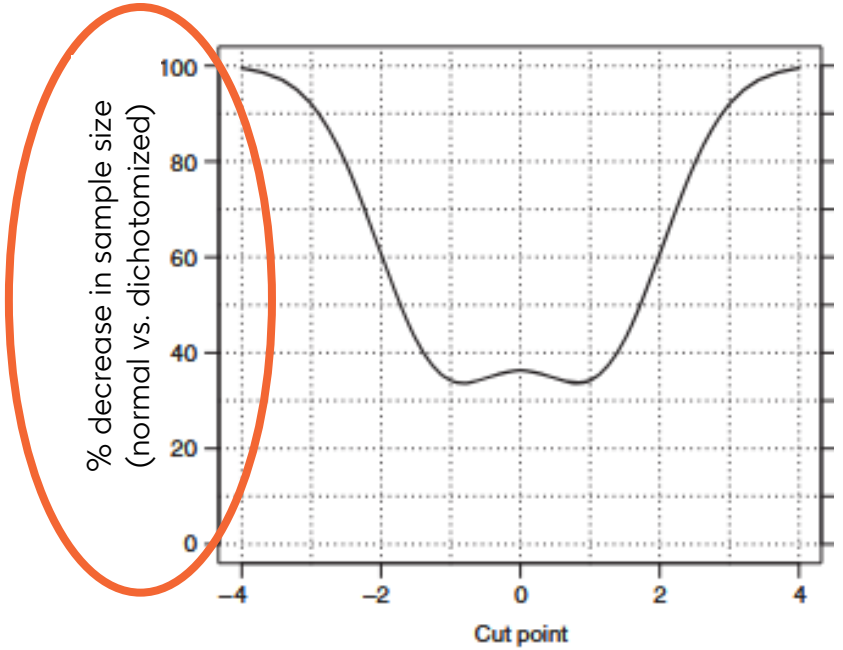
- But despite limitations responder endpoints remain appealing due to
  - Familiarity (with clinicians, patients and regulators)
  - Interpretability
  - Comparability (with previous trials / studies)

- Estimate responder rates by modelling the continuous outcome and computing tail probabilities at the clinically meaningful threshold
  "Keeping clinical interpretability without throwing away information"

GSK

Large efficiency gains when model is correctly specified

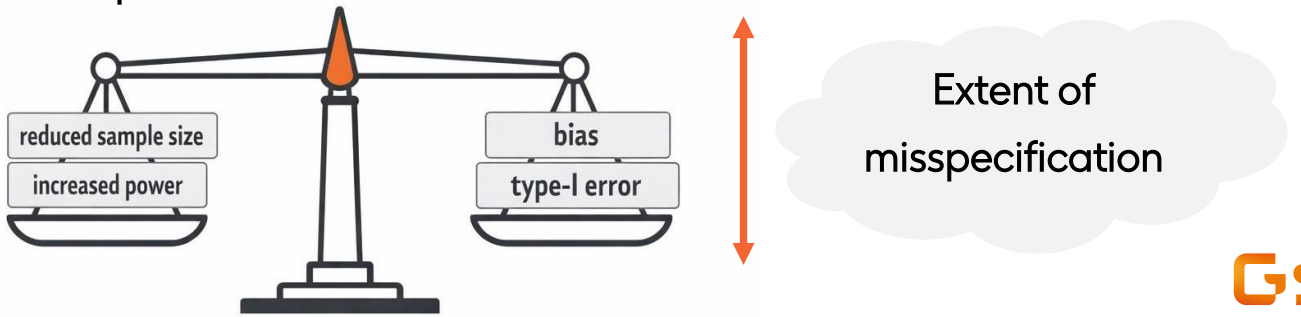
- If the **parametric model is correctly specified**
  - Consistency (asymptotically unbiased)
  - Maximum gain in efficiency
- Sample size improvement vs. dichotomization can be pronounced (AT LEAST 33% SMALLER!)[1]
- However, a **misspecified model** can result in
  - Bias in responder rates and treatment effects
  - Inflated Type-I error
- Critical to check distributional assumptions



Standard Normal Distribution

1. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. *Pharmaceutical Statistics.* 2009;8(1):50-61. doi:10.1002/pst.331

"Bias-Variance tradeoff"



Core Idea

Estimand of interest:

$$p(Y < c \,|T = t)$$

for endpoint Y, responder threshold c and treatment arm T

Parametric Estimator:

1. **Model** the distribution $(Y|X;\theta)$ separately by treatment arm
2. **Predict** response rate for each individual under each treatment $\hat{p}_{it} = F_t(c; X_i, \hat{\theta}_t)$
3. Obtain **marginal response rate** for each arm: $\hat{p}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{it}$ using G-computation
4. Obtain **marginal treatment effect** of interest (risk difference, odds ratio, risk ratio)
5. **Calculate standard errors** via delta method or nonparametric bootstrap

GSK

Methods investigated to protect against misspecification

- We have assessed several modelling choices, all **stratified by arm**:

Model	Details	Key assumption
Logistic regression	Benchmark dichotomized analysis	Binary indicator; binomial inference

Methods investigated to protect against misspecification

- We have assessed several modelling choices, all **stratified by arm**:

Model	Details	Key assumption
Logistic regression	Benchmark dichotomized analysis	Binary indicator; binomial inference
Linear regression	Linear regression	Normal residuals

Methods investigated to protect against misspecification

- We have assessed several modelling choices, all **stratified by arm**:

Model	Details	Key assumption
Logistic regression	Benchmark dichotomized analysis	Binary indicator; binomial inference
Linear regression	Linear regression	Normal residuals
Quantile-Normal (QN) transformed linear regression	Transform full outcome data $Y^* = \Phi^{-1}(\hat{F}(Y))$ Then linear regression	Normal residuals after rank-based transformation

Methods investigated to protect against misspecification

- We have assessed several modelling choices, all **stratified by arm**:

Model	Details	Key assumption
Logistic regression	Benchmark dichotomized analysis	Binary indicator; binomial inference
Linear regression	Linear regression	Normal residuals
Quantile-Normal (QN) transformed linear regression	Transform full outcome data $Y^* = \Phi^{-1}(\hat{F}(Y))$ Then linear regression	Normal residuals after rank-based transformation
Yeo-Johnson (YJ) transformed linear regression	Transform full outcome data (extension of Box-Cox) Then linear regression	Normal residuals after Yeo-Johnson transformation

GSK

Methods investigated to protect against misspecification

- We have assessed several modelling choices, all **stratified by arm**:

Model	Details	Key assumption
Logistic regression	Benchmark dichotomized analysis	Binary indicator; binomial inference
Linear regression	Linear regression	Normal residuals
Quantile-Normal (QN) transformed linear regression	Transform full outcome data $Y^* = \Phi^{-1}(\hat{F}(Y))$ Then linear regression	Normal residuals after rank-based transformation
Yeo-Johnson (YJ) transformed linear regression	Transform full outcome data (extension of Box-Cox) Then linear regression	Normal residuals after Yeo-Johnson transformation
Skew-t regression	Skew-t regression with covariates affecting location (4-parameters: location, scale, skewness, kurtosis)	Skew-t distributed residuals

GSK

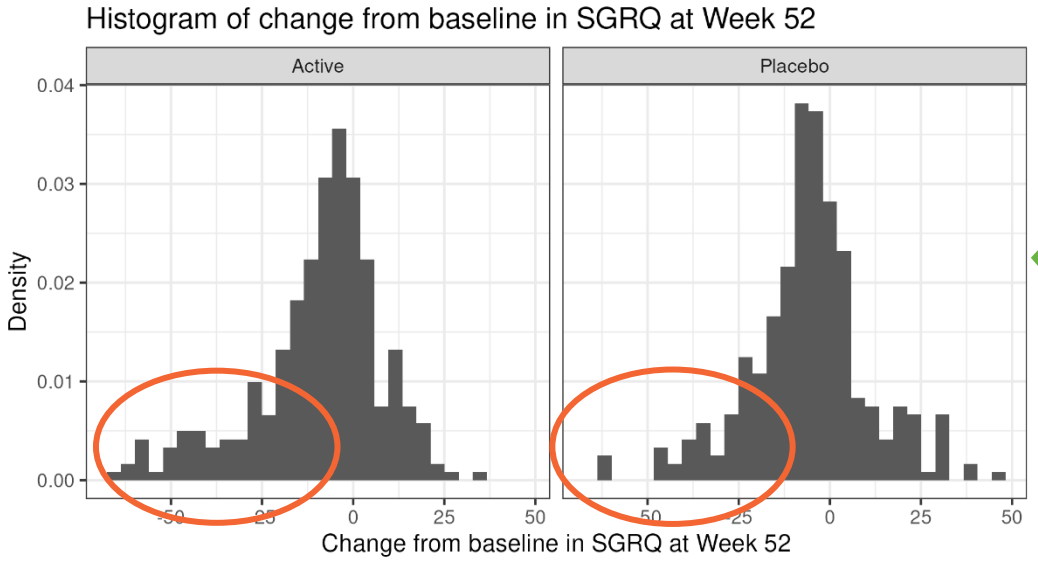
Case Study: PROs in COPD

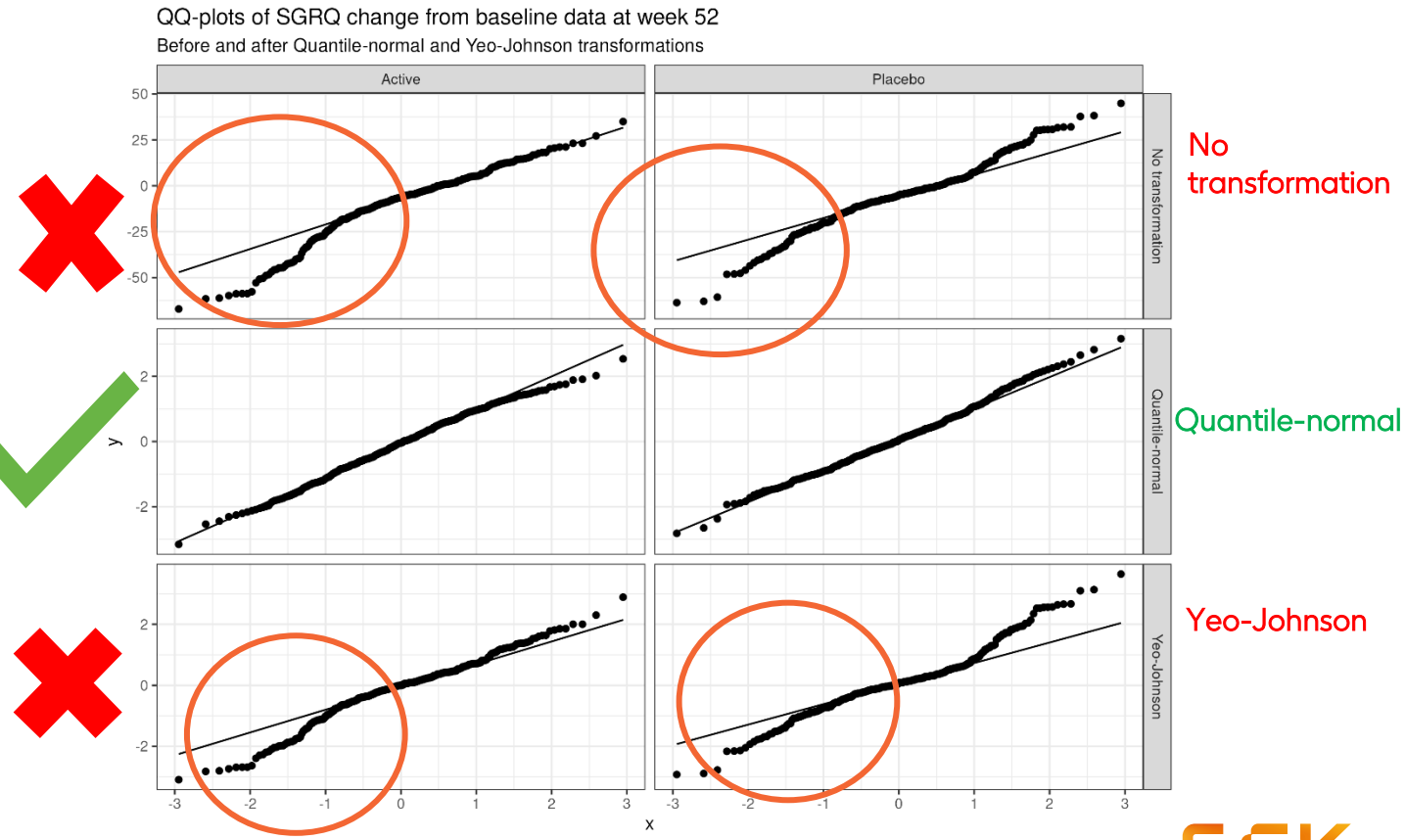
- Patient Reported Outcomes in COPD studies (CAT, SGRQ, E-RS: COPD)
- PRO data from a Phase 3 COPD trial (N= ~800); 1:1 active vs. placebo
- Responder analyses based on a pre-determined improvement from baseline
  - SGRQ Total Score reduction of ≥ 4 points → Responder
  - E-RS: COPD Total Score reduction of ≥ 2 points → Responder
- We analysed the PROs using dichotomized and non-dichotomized transformation-based responder methods.

GSK

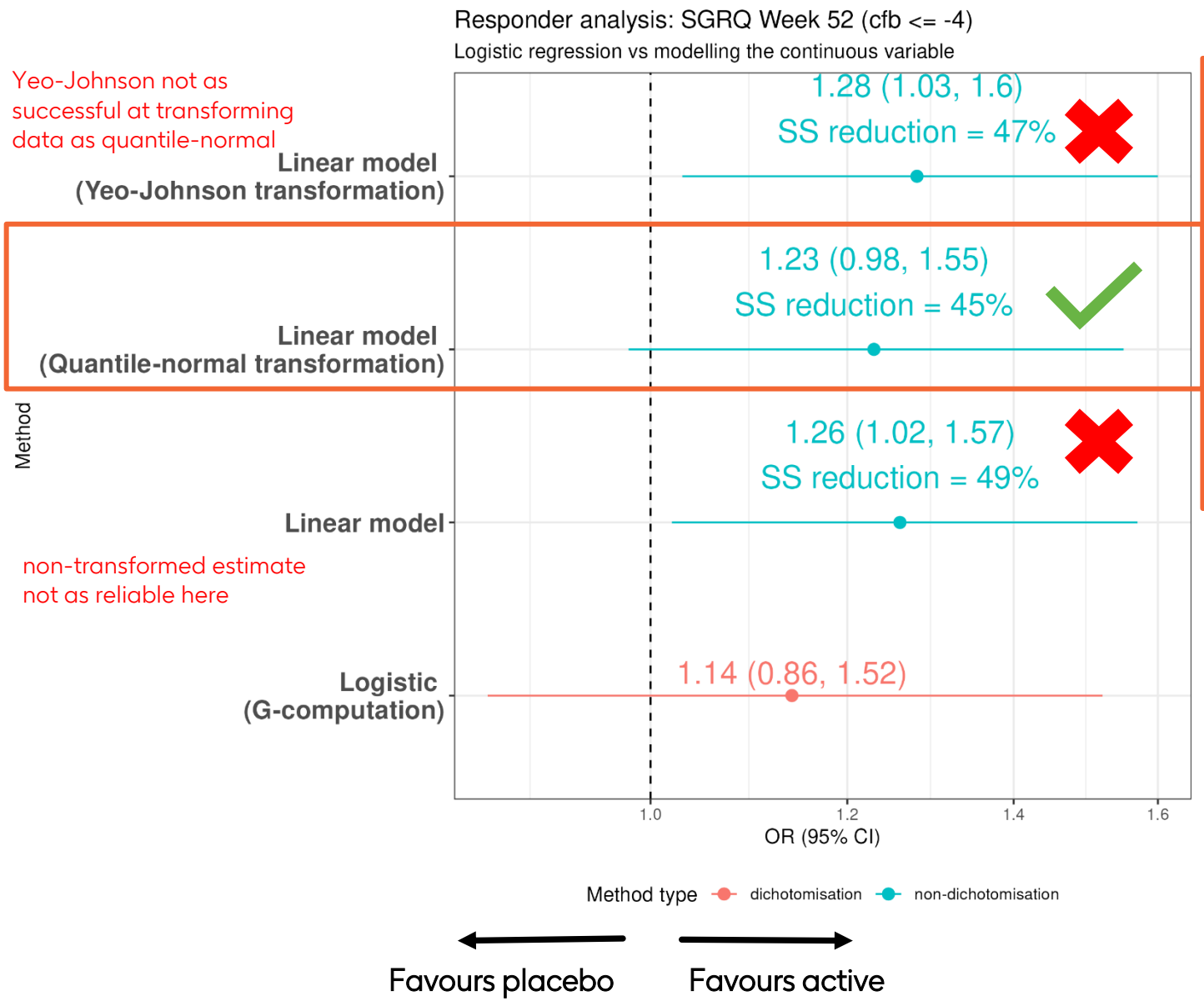
Handling deviations from the normality assumption

- Assessing Normality with diagnostic plots
- For SGRQ – there is some non-normality. Only quantile-normal reliably transformed the data here





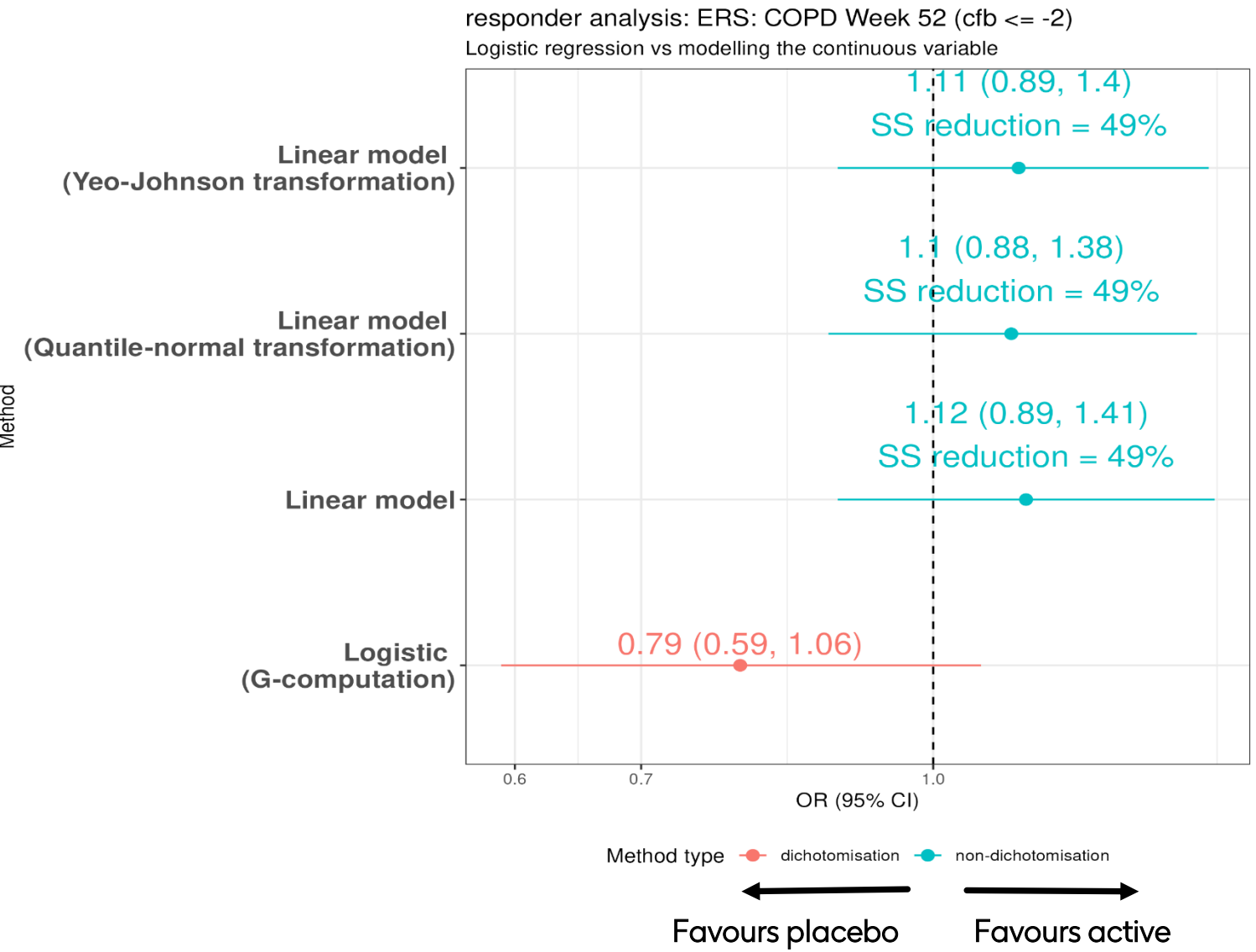
SGRQ responders



- ✓ Non-dichotomisation demonstrates substantial precision gain (45% reduction in sample size)
- ✓ Point estimates comparable to dichotomisation approach

E-RS: COPD results

Discrepancy between dichotomisation and non-dichotomisation approaches



Original results:

- Dichotomized responder analysis point estimate favoured placebo
- Continuous analysis (MMRM) point estimate favoured treatment:

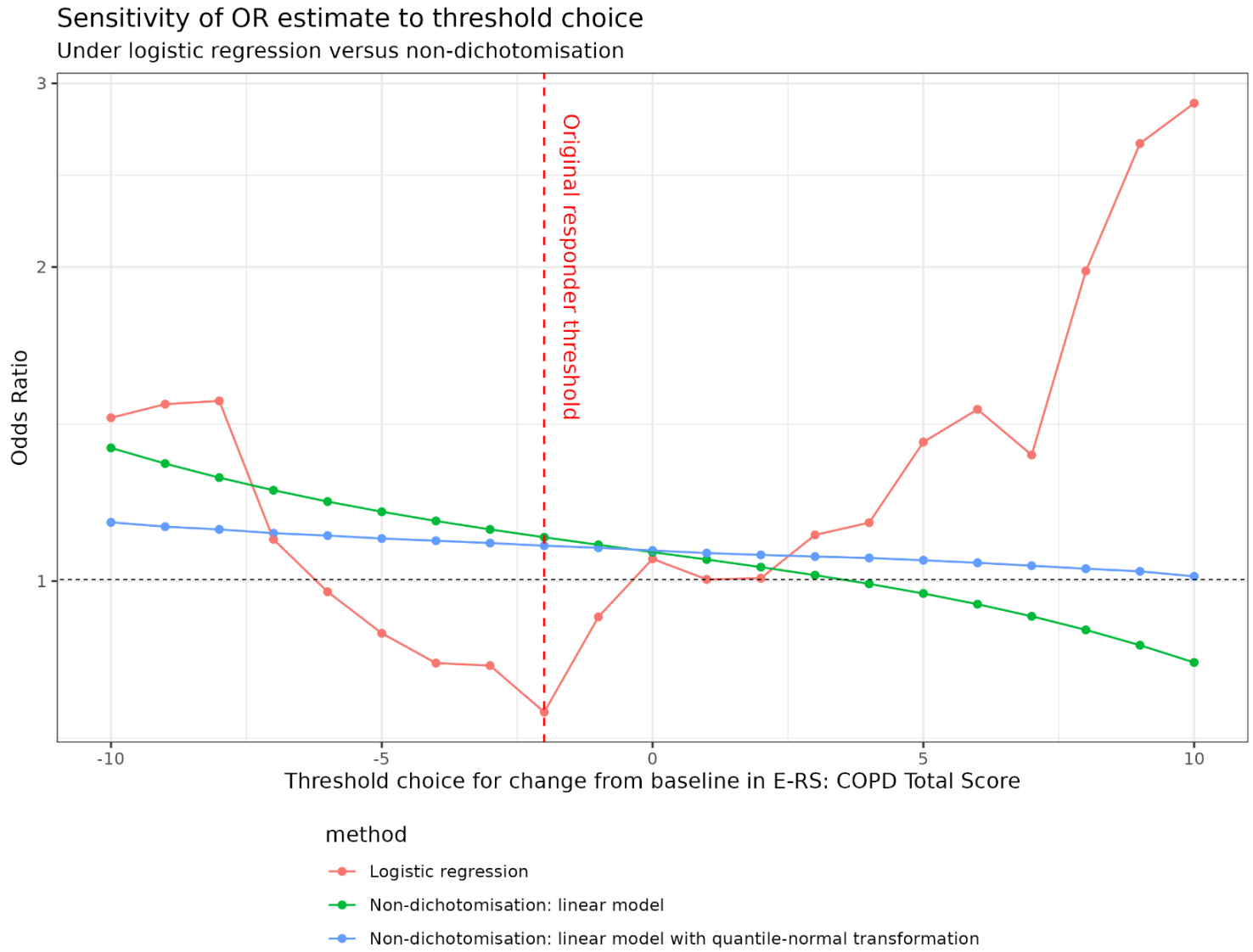
  Active vs Placebo mean difference:

  -0.2 (-1.00, 0.55)

- Non-dichotomized methods aligned with MMRM analysis

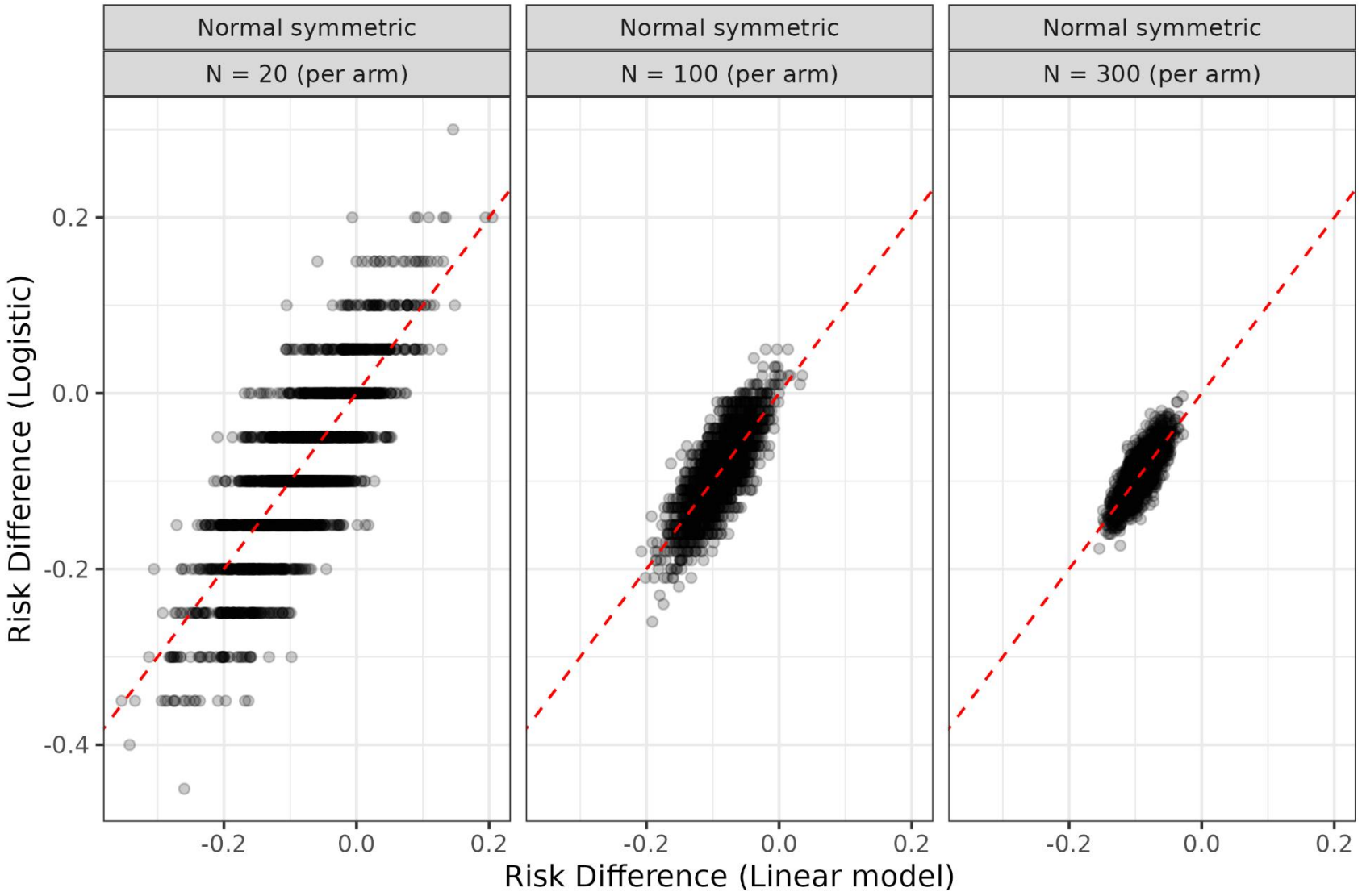
Dichotomized logistic ORs are highly sensitive to response thresholds

ORs estimated using non-dichotomized methods are more stable across response thresholds



Finite-sample differences even under correct specification

Linear model under Normal DGM



Finite-sample differences

- ✓ Between model-based (linear) and empirical (logistic) estimator
- ✓ Even under correct model specification
- ✓ Up to +/- 0.1 on RD scale for N=100 per arm

Handling intercurrent events / missing data

- The following often result in a non-responder classification:
  - Patients experiencing an intercurrent event handled using a composite strategy
  - Patients whose outcome cannot be ascertained due to missing data
- Non-dichotomisation methods require additional modelling in this setting

$$P(\text{Responder}) = P(\text{Responder}|\text{Evaluable})P(\text{Evaluable})$$

Non-dichotomization method on evaluable pts

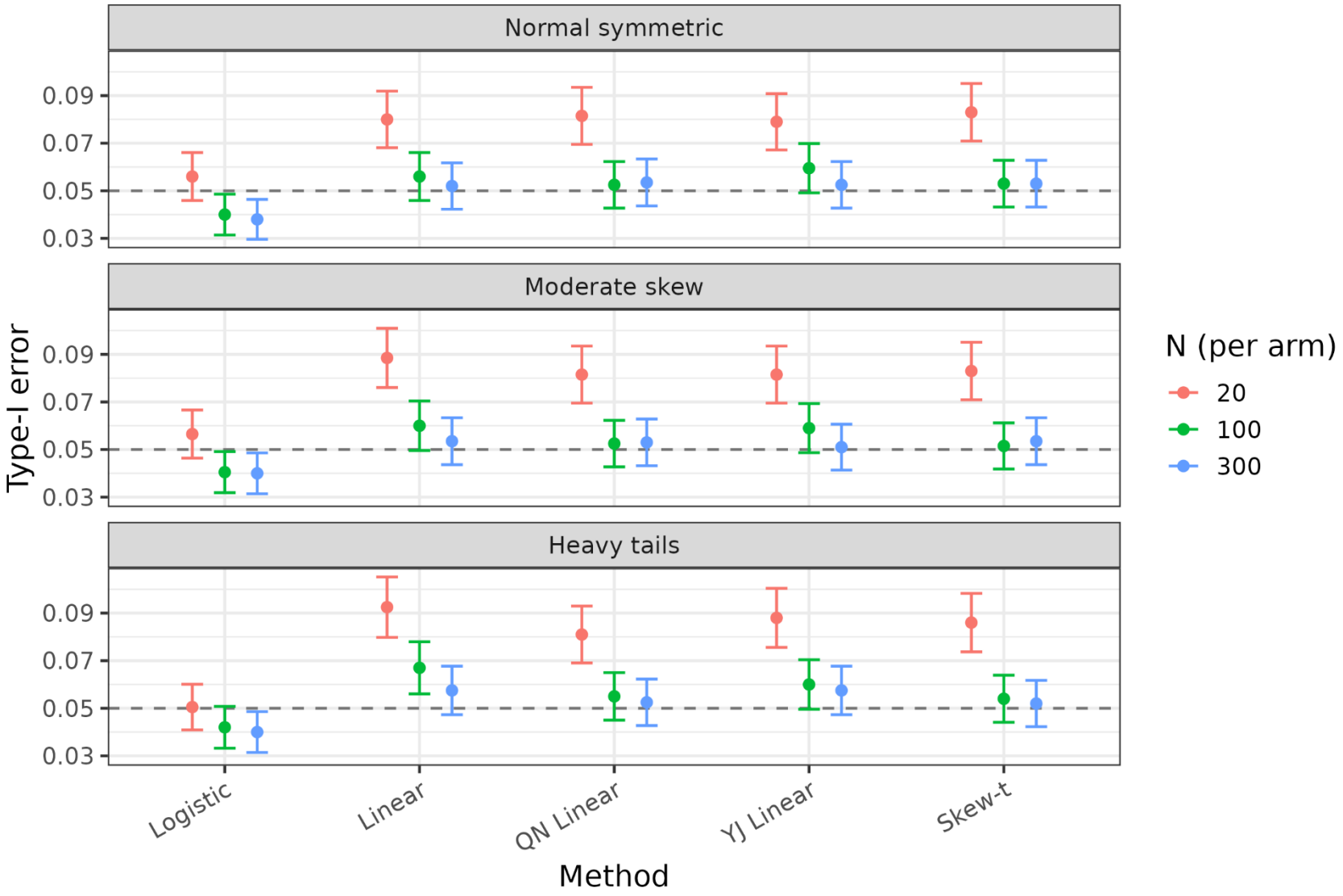
Model (e.g. logistic)

where "Evaluable" data = No composite ICE and no missing data

- Robust approach but efficiency decreases as non-evaluable pts increase

GSK

Type-I error rate by sample size (simulation study results)



Bootstrapped Type-I error

- ✓ Type-I error inflated for small “P2-type” trials
- ✓ For N > 100 most models control type-I even under misspecification

Conclusions

- Non-dichotomization methods can lead to substantial increase in efficiency
- However model misspecification can cause bias
- Composite ICEs / missing data require additional modelling but remain naturally accommodated within the framework
- Type-I error reasonably controlled for sample sizes >100 if bootstrap SEs/CIs used
- Differences are likely between empirical and model-based estimates
- Early regulatory engagement critical for justifying model-based responder analyses.
- Practical Recommendation:

QN-linear + G-computation + bootstrap SE/CI, with prespecified diagnostics/sensitivity; consider skew-t when heavy tails /skew suspected.

GSK

gsk.com

Simulation Study: Data Generating Mechanisms

3 scenarios:

1. Normal Distribution
2. Moderate Negative Skew (same for both arms)
3. Skewness with Heavy Tails (same for both arms)

Null and alternative hypotheses

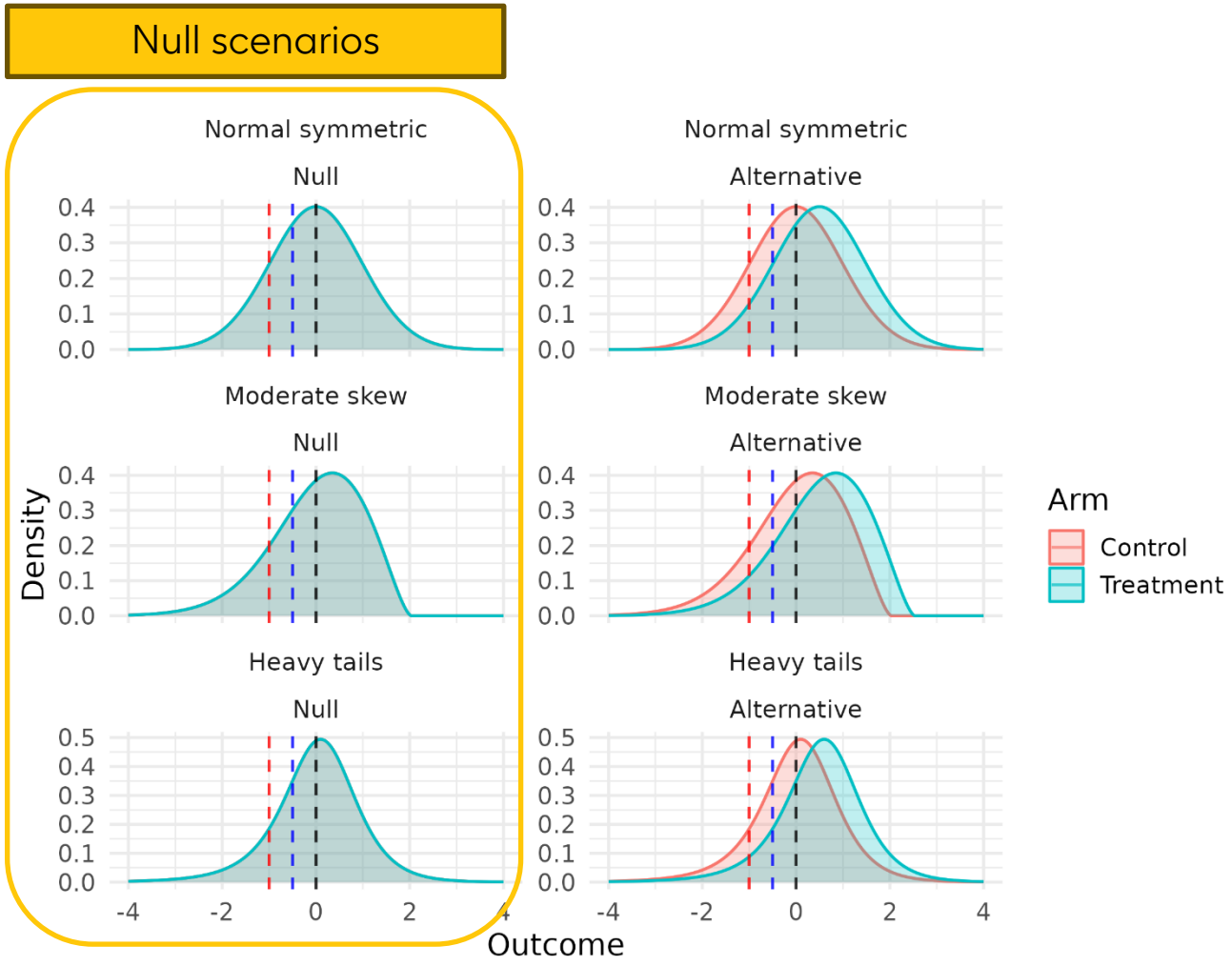
- **Null:** Zero location shift (treatment vs. control)
- **Alternative:** Treatment location shift = + 0.5

3 responder thresholds

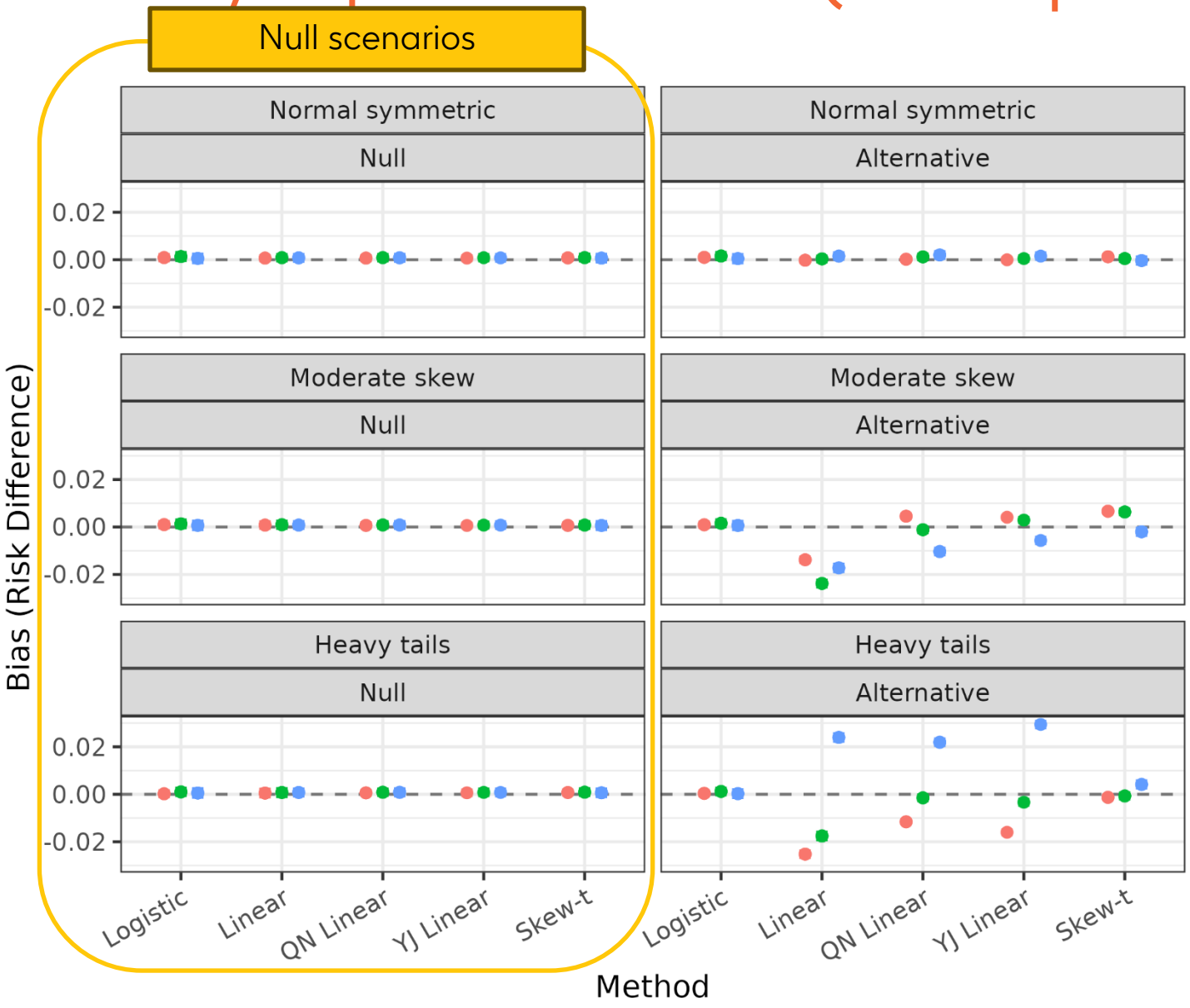
- Responder if ≤ -1, ≤ -0.5, ≤ 0

Sample Sizes

- 20, 100 & 300 per arm,



Bias by responder threshold (N=300 per arm)

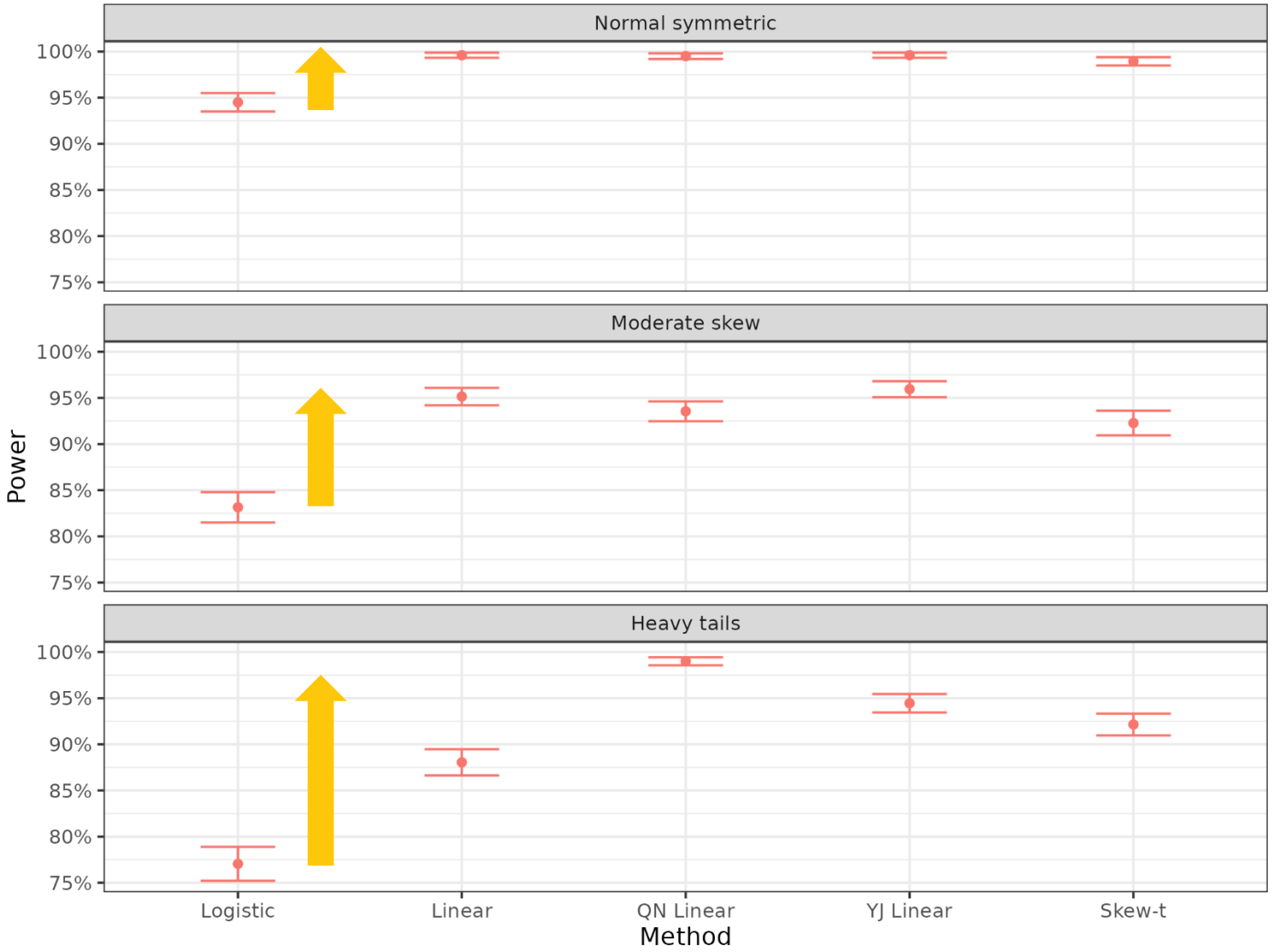


- ✓ Little to no bias under null and Normal scenarios
- ✓ Linear method biased under non-normality
- ✓ Q-N and Y-J transform biased under heavy tails
- ✓ Skew-t performs reasonably well

2000 simulations

GSK

Power at threshold = -1 , N = 300 per arm



Power

- ✓ Substantial increase in power compared to logistic
- ✓ Results need to be interpreted in combination with bias

Example: Linear Regression

1. **Model:** $Y_i \sim N(\beta_t^T X_i, \sigma_t)$ for each treatment arm t separately
2. **Predict:** $\hat{p}_{it} = \Phi\left(\frac{c - \hat{\beta}_t^T X_i}{\hat{\sigma}_t}\right)$ from tail area of Normal distribution evaluated at c
3. **G-computation:** $\hat{p}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{it}$
4. **Marginal Odds Ratio:** $\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$
5. **Calculate standard errors** via delta method or nonparametric bootstrap