Bootstrap Methods for Confidence Interval Estimation in Small Samples: Implications for Health Economics

João Rocha[1] | Jorge Félix[1]

1. Exigo Consultores, Lisbon, Portugal

MSR48

OBJECTIVES

In small-sample settings or when data are skewed – common situations in health economics – the assumptions required for parametric methods to perform well are often violated. As a result, confidence intervals derived from these methods may be inaccurate and exhibit poor coverage of the parameter of interest.

This study aims to explore, through a simulation, non-parametric bootstrap techniques as alternatives for confidence interval estimation, given small samples, in the context of Health Technology Assessment (HTA), and highlight the limitations of parametric approaches.

METHODS

A Monte Carlo simulation was conducted to evaluate three approaches for estimating the pooled log odds ratio (logOR) and corresponding confidence intervals (with 95% confidence) in rare-event meta-analyses: 1) a Generalized Linear Mixed Model (GLMM) with profile-likelihood confidence intervals; 2) GLMM with non-parametric bootstrap – Percentile and Bias Corrected accelerated (BCa); 3) Bayesian random-effects model with weak informative priors – N(0,2.5) for treatment effect and Exp(5) for between-study standard deviation (SD).

Classical parametric approaches (e.g., Wald-based confidence intervals and Mantel–Haenszel pooling) were not included, as they have been largely abandoned in current HTA practice due to their poor performance in sparse data and small-sample contexts [1].

Synthetic datasets were generated to mimic binary rare-event outcomes. For each simulated meta-analysis, k studies were created, each with binomially distributed event counts per arm:

$$y_{Ci} \sim Binom(n_{Ci}, \mathrm{p_{Ci}}), \qquad y_{Ti} \sim Binom(n_{Ti}, \mathrm{p_{Ti}})$$

with

$$p_{Ci} = logit^{-1}[logit(p_C) + \delta], \delta \sim N(0,\tau) \quad \text{and} \quad p_{Ti} = p_{Ci} * Relative\,risk$$

Key design parameters were varied following prior simulation frameworks [2], with the number of studies (k) set to 5, 10, and 30, control arm event probabilities (p_C) of 0.01 and 0.05, heterogeneity (τ) values of 0.5 and 1, and a fixed relative risk of 0.5. Each scenario was replicated 100 times, and 2000 bootstrap samples were drawn for each confidence interval.

Performance was evaluated through five different metrics by their average results: Coverage probability; Width; Bias; Power; Model estimability.

RESULTS

For low event probability (1%), bootstrap-based GLMMs showed lower coverage (~80%) than profile-likelihood and Bayesian approaches, which were close to the nominal 95% across all settings (**Figure 1**).

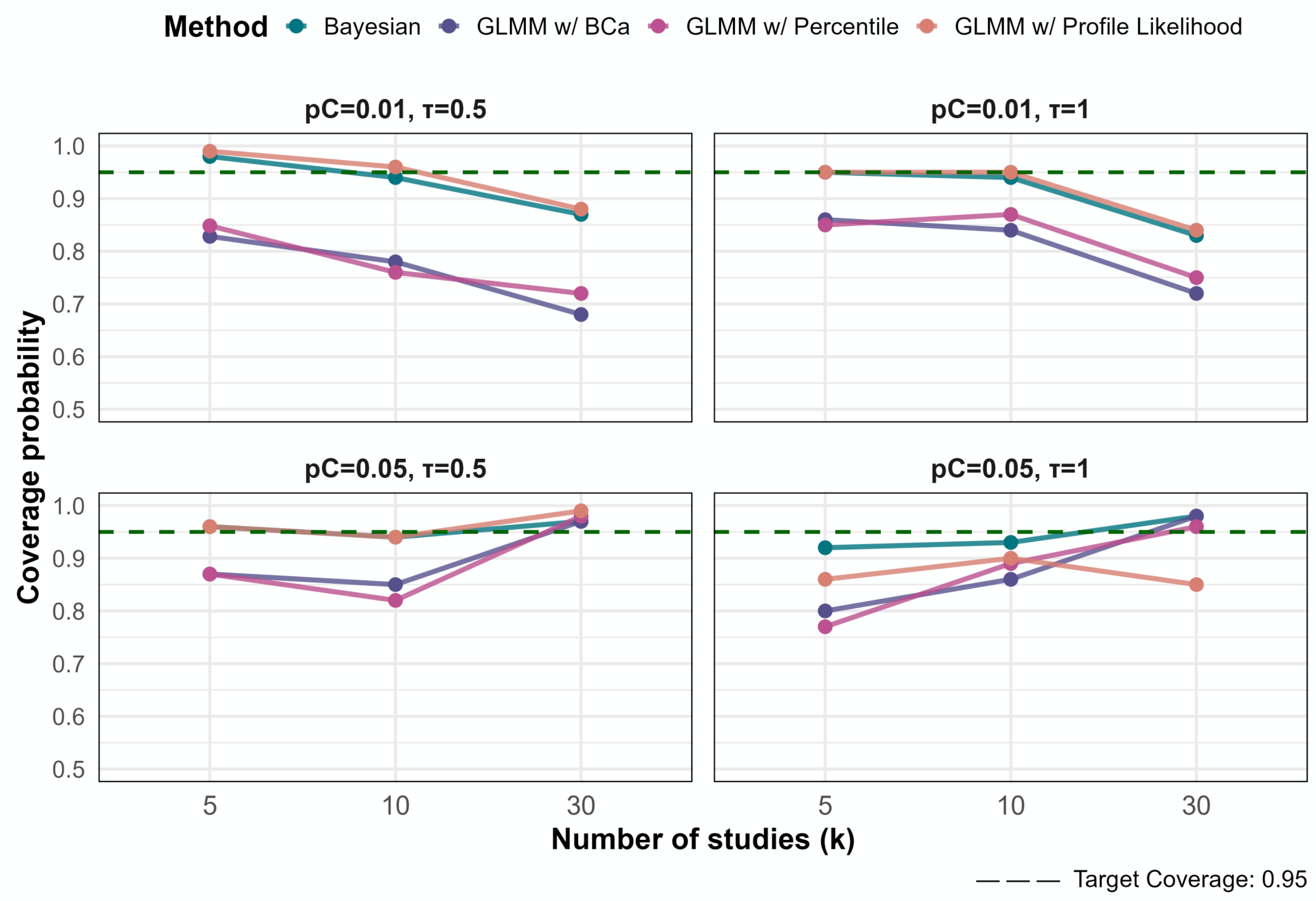

Figure 1. Coverage probability average results.

RESULTS (cont.)

For higher event probability (5%) and several studies (k = 30), bootstrap methods also achieved coverage close to 95% (**Figure 1**).

For a small number of studies (k = 5), bootstrap methods yielded confidence intervals two to four times wider than those from profile-likelihood and Bayesian approaches, regardless of event probability or heterogeneity (**Figure 2**).

However, with a moderate number of studies (k = 10), the interval widths became comparable across methods (**Figure 2**).

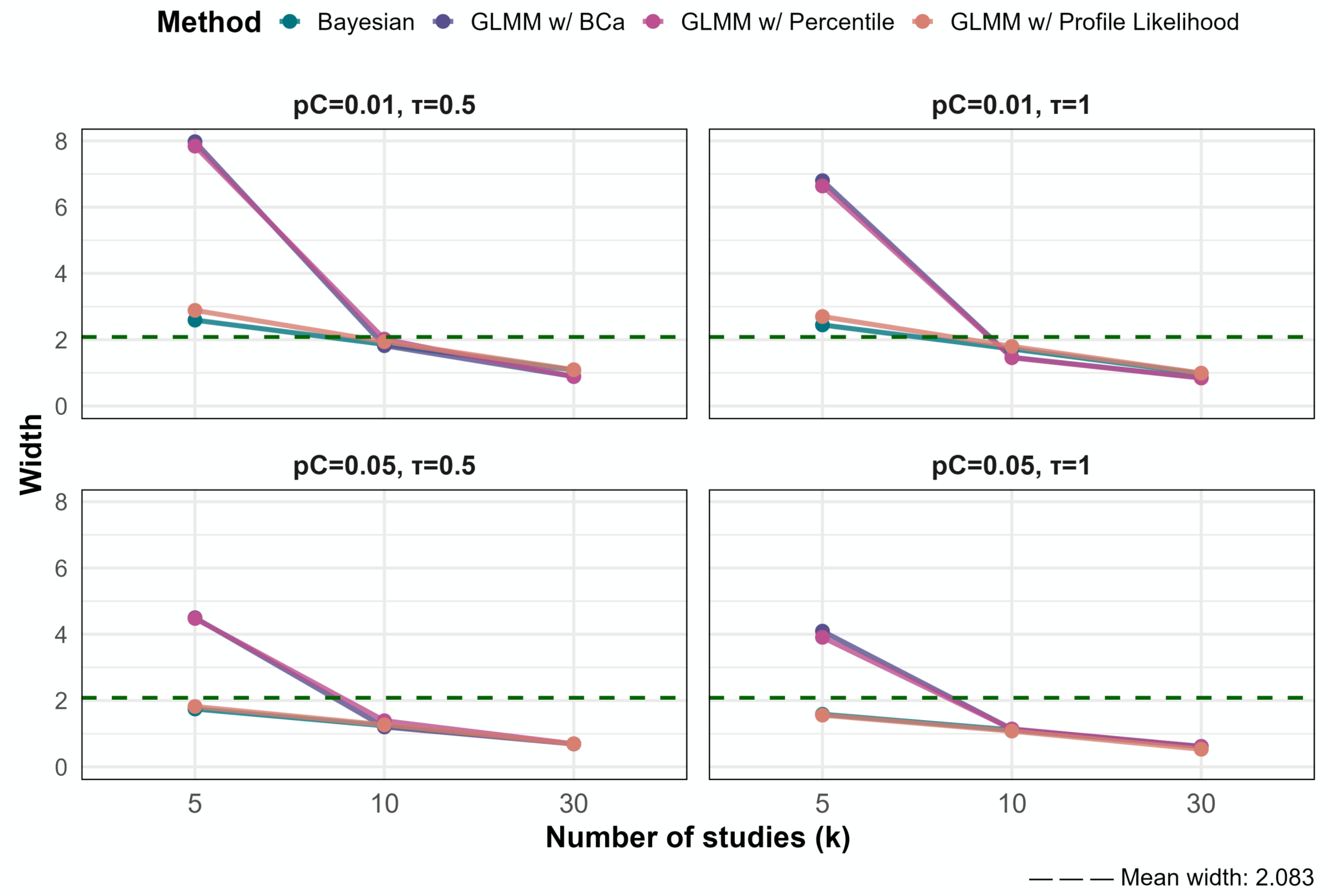

Figure 2. Width average results.

A composite performance score was developed to identify the best-performing method. Each metric was normalized (min-max) and weighted: coverage and interval width (0.4 each); bias and power (0.1 each); estimability was excluded as all models converged.

For low event probability in control arm (1%), bootstrap methods performed 1.25 – 2 times worse than the other approaches, regardless of the number of studies. However, at a 5% event probability and with 30 studies, their performance became comparable (scores ≈ 0.8), as shown in **Figure 3**.

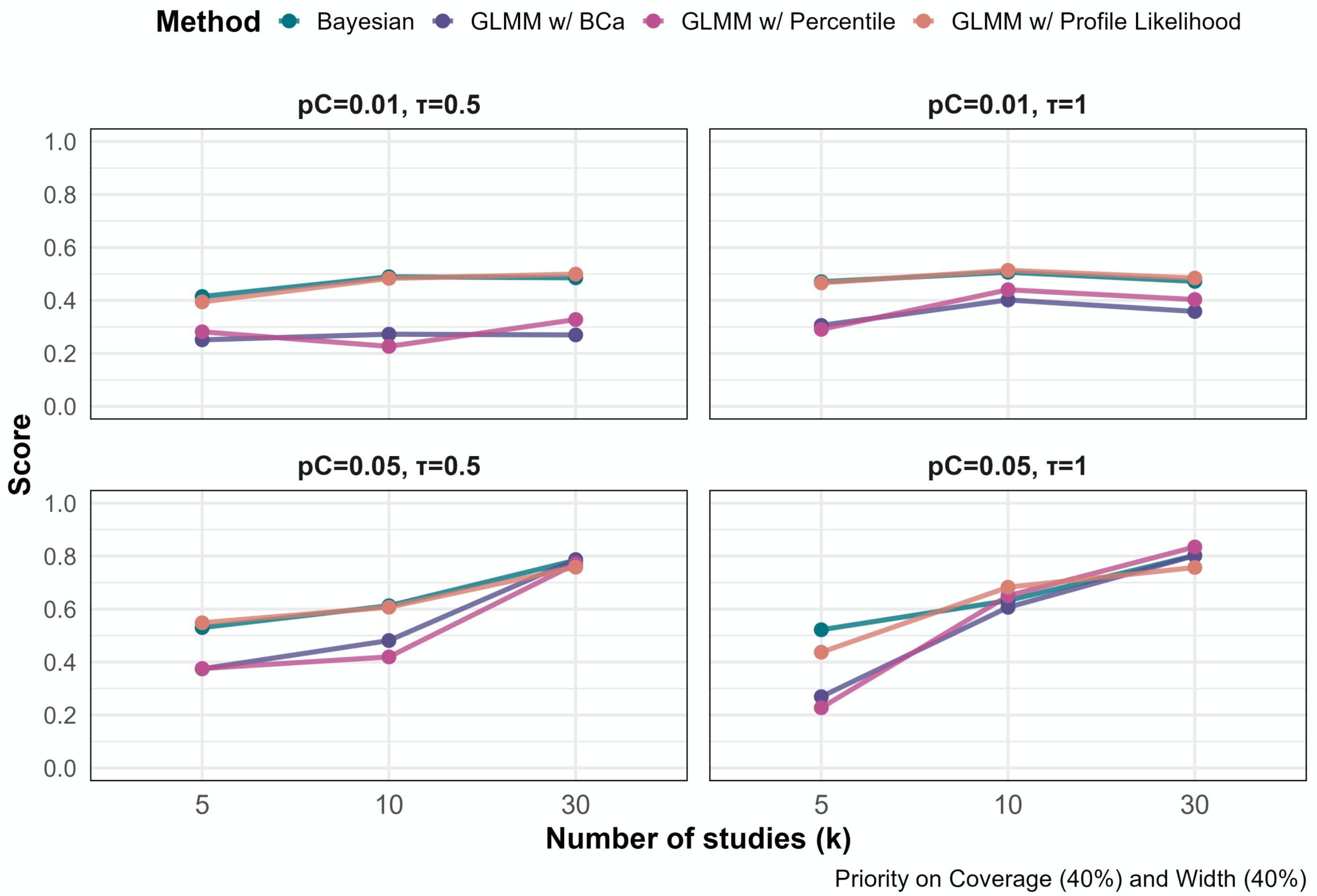

Figure 3. Score average results.

=CONCLUSIONS

Bootstrap methods are a viable alternative to classical parametric approaches to calculate confidence intervals in meta-analyses. They perform has good as the standard methods in settings with a great number of studies (~30) and a moderate event incidence (~5%). However, in settings with fewer studies or lower event incidence, GLMM with profile likelihood and Bayesian models provide a more reliable inference.

[1] Hodkinson A, Kontopantelis E. *Applications of simple and accessible methods for meta-analysis involving rare events: A simulation study*. Stat Methods Med Res. 2021 Jul;30(7):1589-1608.
[2] Yao, M., Deng, K., Wang, Y. et al. *Random-effects meta-analysis models for pooling rare events data: a comparison between frequentist and bayesian methods*. BMC Med Res Methodol 25, 228 (2025).

ISPOR Europe 2025 | November 2025 | Glasgow, Scotland, UK