Improving Causal Inference from Observational Data

A Comparative Analysis of Confidence Interval Methods in Sequential Target Trial Emulation

Florian Metwaly

Background

Sequential Target Trial Emulation (TTE) is a framework that is gaining popularity to estimate causal effects from observational data when a randomized control trial (RCT) is not feasible. As sequential TTE involves copying data from the observational dataset multiple times, the assumption of independence between observations is violated, and confidence intervals (CI) can no longer be estimated by standard estimation approaches. While sandwich-type estimators are commonly implemented in statistical software for CI estimation, the non-parametric bootstrap is the recommended method for CI estimation in the TTE literature. We compare the performance of the two approaches and present the Julia package *TargetTrialEmulation.jl*, for computationally efficient estimation of bootstrap CIs in the context of sequential TTE.

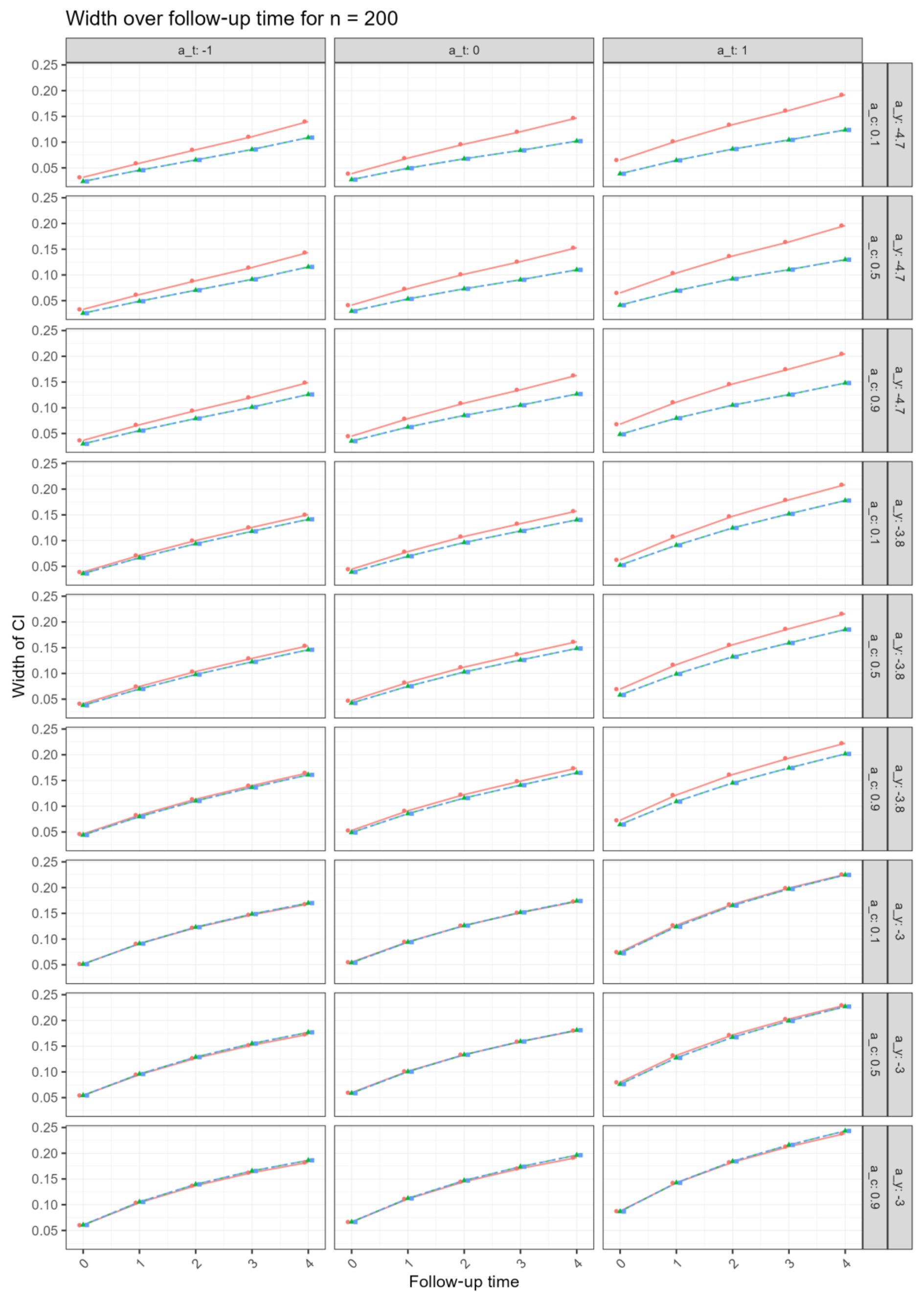

Figure 1

Width of the 95% confidence interval over 1000 simulations with sample size n = 200. The green line denotes the empirical bootstrap CI, the blue line denotes percentile bootstrap CI, and the red line denotes sandwich estimator CI. The outcome event rate indicator is denoted by a_y, the confounding strength is denoted by a_c, and the treatment prevalence indicator is denoted by a_t.

Methods

In a simulation study, we compare the performance of the sandwich-type against two types of non-parametric bootstrap CI estimation: the percentile and the empirical bootstrap. We test the estimators across 81 scenarios, varying the sample sizes, confounding strength, outcome prevalence, and treatment prevalence. We compare their performance regarding their coverage, width, bias-eliminated coverage, and power.

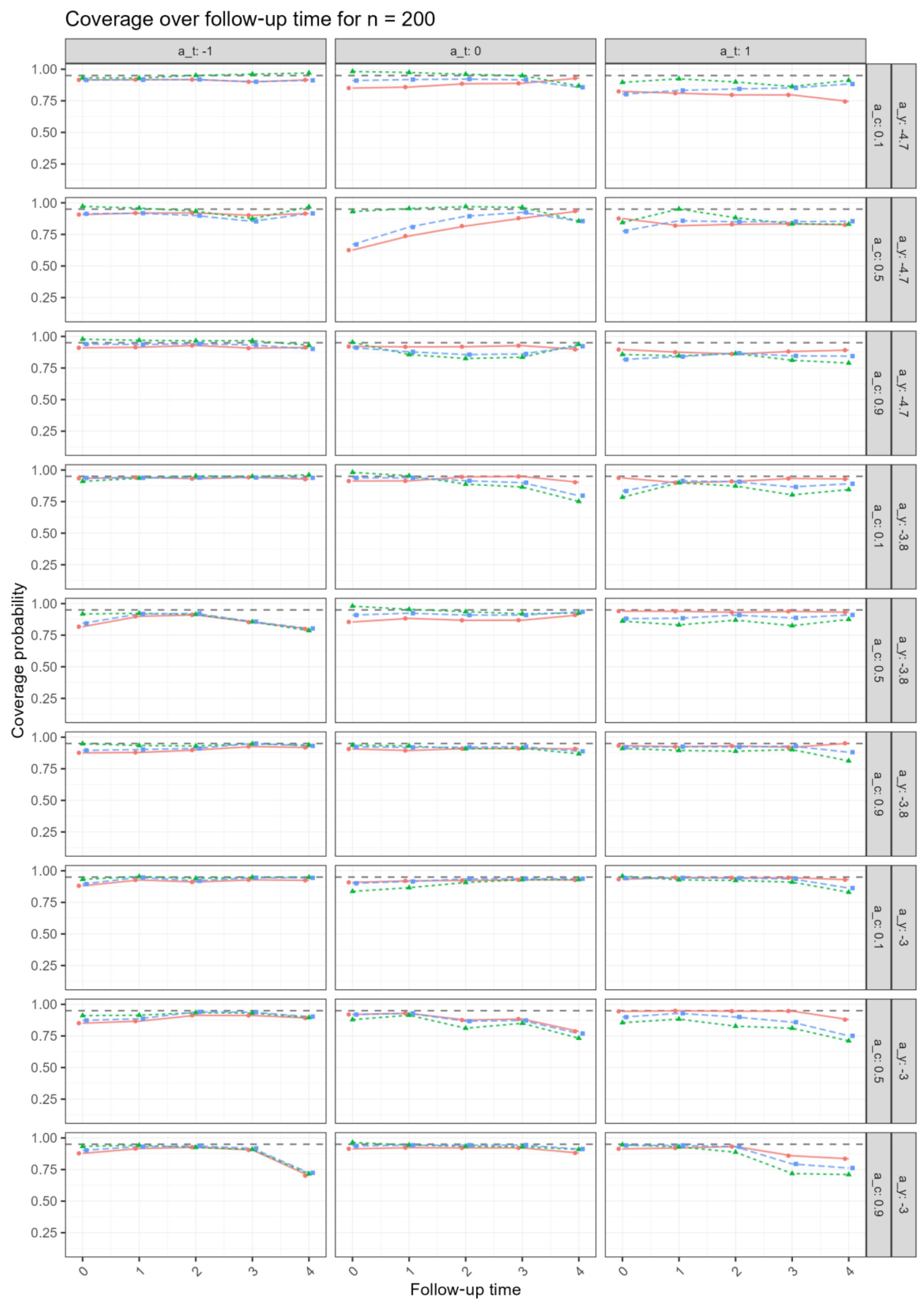

Figure 2

Coverage of the 95% confidence interval over 1000 simulations with sample size n = 200. The green line denotes the empirical bootstrap CI, the blue line denotes percentile bootstrap CI, and the red line denotes sandwich estimator CI. The dashed gray line marks 95%. The outcome event rate indicator is denoted by a_y, the confounding strength is denoted by a_c, and the treatment prevalence indicator is denoted by a_t.

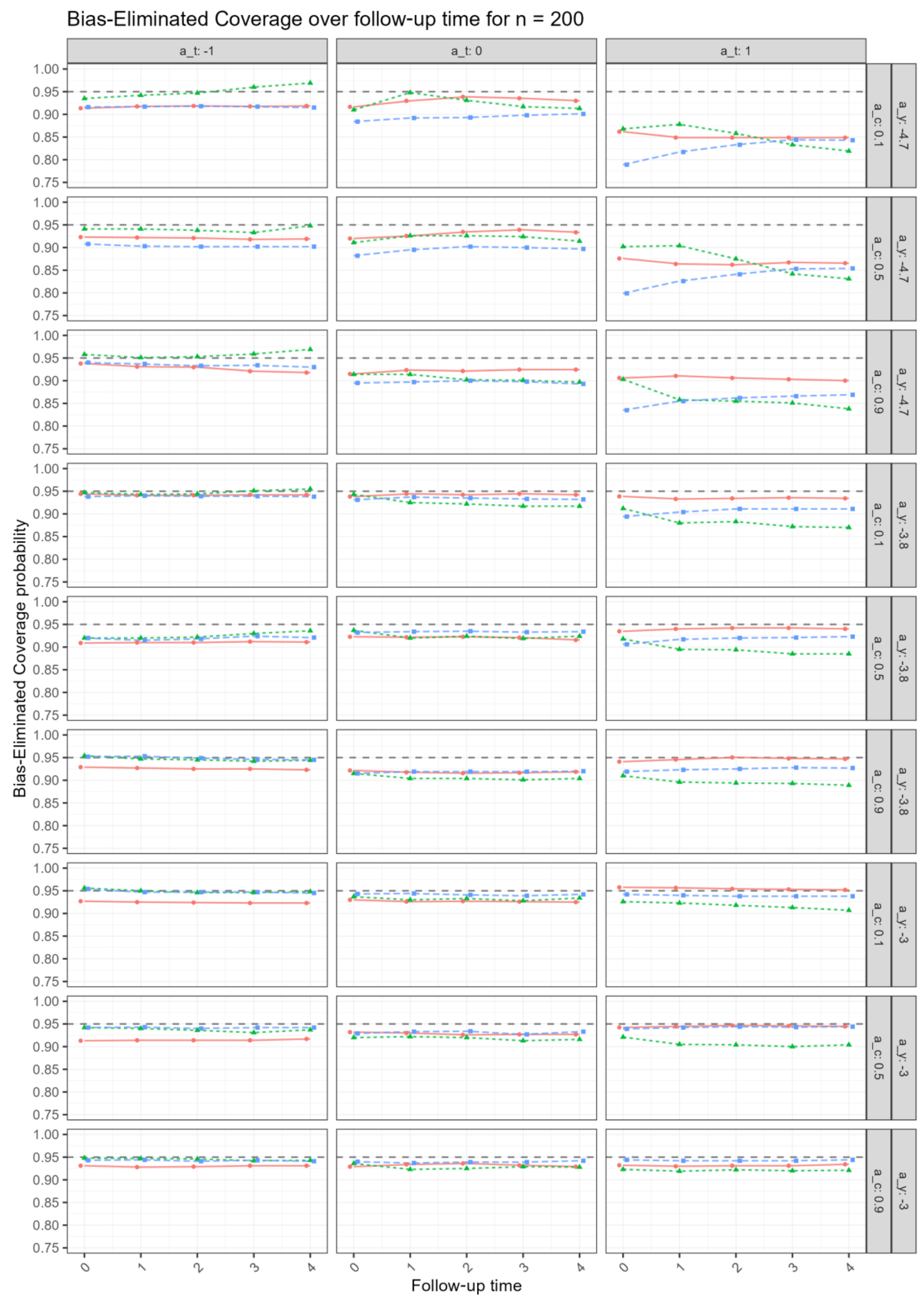

Figure 3

Bias-eliminated coverage of the 95% confidence interval over 1000 simulations with sample size n = 200. The green line denotes the empirical bootstrap CI, the blue line denotes percentile bootstrap CI, and the red line denotes sandwich estimator CI. The dashed gray line marks 95%. The outcome event rate indicator is denoted by a_y, the confounding strength is denoted by a_c, and the treatment prevalence indicator is denoted by a_t.

Results

Our simulation study finds that both bootstrap methods, percentile and empirical, more frequently achieved coverage closer to the nominal 95% level, particularly in small sample sizes, and produced narrower confidence intervals than the sandwich estimator. Although the sandwich-type estimator found consistently higher statistical power, this finding is most likely attributable to the sandwich-type estimator being biased upwards. Bias-eliminated coverage analysis suggested that coverage differences were primarily driven by bias in the point estimates, particularly at low event rates and later follow-up times.

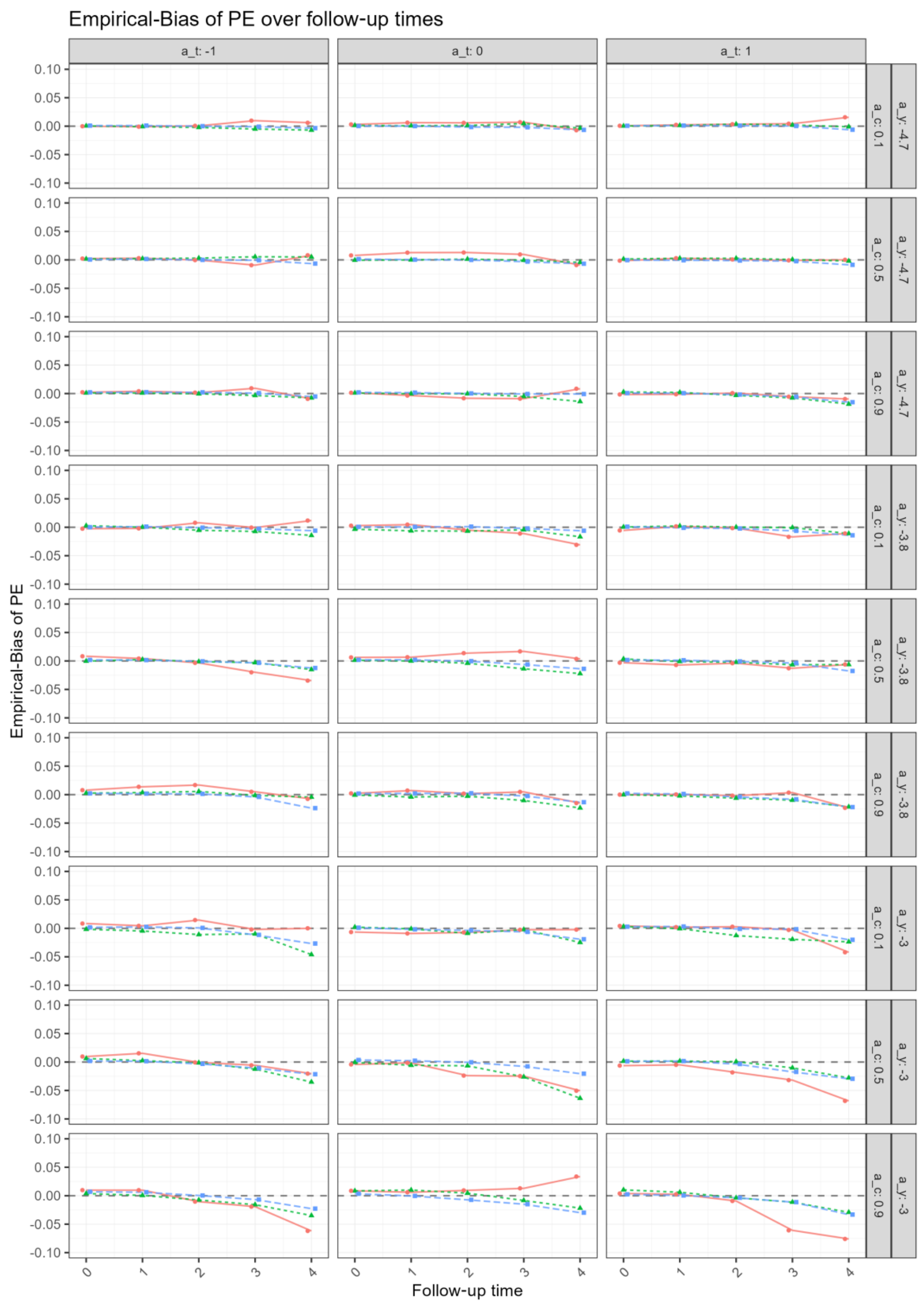

Figure 4

Bias of the MRD point estimates over 1000 simulations. The red line denotes the sample size n = 200, the green line denotes the sample size n = 1000, and the blue line denotes the sample size n = 5000. The dashed gray line marks the reference value 0. The outcome event rate indicator is denoted by a_y, the confounding strength is denoted by a_c, and the treatment prevalence indicator is denoted by a_t.

Conclusion

While this study confirms bootstrap's potential, limitations of the percentile and empirical approaches, such as sensitivity to skewed distributions and a biased estimator, suggest the need for more robust methods. Advanced techniques like the bias-corrected and accelerated (BCa) bootstrap may provide improved inference in complex TTE settings. Future work should focus on benchmarking these methods using the evaluation framework employed in this study, coverage, width, and bias-eliminated coverage. It should additionally incorporate assessments of power and type I error control, and further focus on computational efficiency.

f.j.metwaly@uu.nl

https://github.com/flo1met/thesis_TTE