MSR84

# Empirical Comparison of Four Approaches for Generating Confidence Intervals Around Weighted Trial-Level Correlation

**Paul Serafini, Victoria Wan, Mir Sohail Fazeli, Murat Kurt**
Evidinno Outcomes Research Inc., Vancouver, BC, Canada

## Background

- A surrogate endpoint is an intermediate outcome which substitutes for or predicts a final "true" outcome of interest in a clinical trial.[1]
- Surrogate endpoints are increasingly used in clinical trial designs to accelerate the assessment of new therapies when collecting statistically mature data for the true endpoint is time-consuming, logistically complex, or costly.[2]
- Elston and Taylor proposed three criteria for the validation of a surrogate endpoint: biological plausibility of a causal mechanism between the surrogate and true endpoints, association between the surrogate and true endpoints at the individual patient level, and association between treatment effects on the surrogate and true endpoints across clinical trials.[1]
- Trial-level association (i.e., association between the treatment effects on the surrogate and true endpoints across trials) is often quantified by the weighted Pearson's correlation and its confidence interval (CI), weighted by sample size or the inverse of variances of the treatment effect estimates for the surrogate or true endpoint. Per Fisher (1925),[3] the standard error (SE) for a Fisher transformed Pearson's correlation is approximately expressed by

$$\sqrt{1/(n-3)}$$

  where *n* is the number of pairs of observations (in surrogacy, the number of trials). However, the appropriate choice of *n* in this formula is ambiguous in the case of a weighted correlation as the trials in the data set do not contribute equally to the correlation coefficient. As a consequence, researchers often use bootstrapping to derive a CI.
- In bootstrapping, the sample data are repeatedly sampled with replacement, and one estimate ("replicate") is calculated for each sample. The resulting distribution of replicates is then used to approximate the sampling distribution of the statistic.
- An ideal procedure for the generation of a CI should capture the underlying value of the parameter of interest at approximately the specified confidence rate across repeated samples. For example, a method for calculating a 95% CI should capture the underlying parameter of interest across 95% of samples.
- To our knowledge, the performance of different methods for calculating a CI on the weighted Pearson's correlation have not been assessed in the literature.

## Objectives

- Primary: To compare the performance of four alternative methods for calculating a 95% CI on the weighted Pearson's correlation using synthetic data.
- Secondary: To investigate the stability of the results across alternative approaches with respect to key parameters of the CI generation process.

## Methods

### Approaches for Confidence Interval Generation

- Four methods of calculating a 95% CI around the weighted Pearson's correlation were compared: (1) SE-based CI defining *n* as effective sample size (ESS) per Hill (1973),[4] (2) SE-based CI defining *n* as ESS per Kish (1965),[5] (3) SE-based CI defining *n* as the number of studies in the evidence base (the "naïve" method), and (4) bootstrapped CI.
- Per Hill (1973), given *w* is a pre-specified vector of normalized weights for the studies ($w_i$) in the evidence base such that $w_i \geq 0$ and $\sum w_i = 1$, ESS is defined as $n_H$:

$$H = -\sum w_i ln(w_i)$$

$$n_H = \exp(H)$$

- Per Kish (1965), where *w* is a vector of weights for the studies ($w_i$) in the evidence base, which may not necessarily be normalized, ESS is defined as $n_K$:

$$n_K = \frac{(\sum w_i)^2}{\sum w_i^2}$$

- Bootstrapping was conducted with 1,000 replicates, and the 2.5% and 97.5% quantiles of the bootstrapped sampling distribution were used to estimate the 95% CI.

### Simulation

- The four methods were compared in a Monte Carlo simulation implemented in C++ using the Armadillo library to generate bivariate normal (BVN) data.[6]
- A total of 64 experimental settings were explored, representing all combinations of the following parameters: (1) the number of studies (*N*) in the evidence base (10, 20, 40, and 80), (2) between-study correlation $\rho_b$ (0.0, 0.3, 0.6, and 0.9), (3) within-study correlation $\rho_w$ (0.0, 0.3, 0.6, and 0.9).
- In each experiment, the following procedure was repeated 10,000 times:
  1. For each study *i*, a pair of treatment effect parameters on the surrogate and true endpoints ($\mu_{i,x}$ and $\mu_{i,y}$) were drawn from a BVN distribution with marginal means of 0, between-study variances ($\sigma^2_{b,x}$ and $\sigma^2_{b,y}$) of 1, and a between-study correlation of $\rho_b$.
  2. Then, for each study *i*, a pair of sample treatment effects on the surrogate and true endpoints ($x_i$ and $y_i$) were drawn from a BVN distribution with corresponding means of $\mu_{i,x}$ and $\mu_{i,y}$, within-study sampling variances of $\sigma^2_{w,i}$ (drawn from a uniform distribution ranging from 0.1 to 0.9), and a within-study correlation of $\rho_w$.
  3. The sample Pearson's correlation *r* between *x* and *y* (i.e., the treatment effects on the surrogate and true endpoints) w as calculated, weighting each study by its inverse sampling variance, $1/\sigma^2_{w,i}$.
  4. Finally, 95% CIs around *r* were calculated using the four methods.
- For each method, the coverage rate, defined as the fraction of the 10,000 replications in which the computed 95% CIs captured the true correlation, was compared to the pre-specified confidence rate of 95%. The underlying value of the correlation parameter in the simulation can be calculated by the following closed form expression:

$$\frac{\rho_b \sigma_{b,x} \sigma_{b,y} + \rho_w \sigma_{w,x} \sigma_{w,y}}{\sqrt{(\sigma^2_{b,x} + \sigma^2_{w,x})(\sigma^2_{b,y} + \sigma^2_{w,y})}}$$

  Because $\sigma^2_{i,w}$ varied by study, its expected value (0.5, the center of the uniform distribution ranging from 0.1 to 0.9) was used in this calculation for the study-independent within-study sampling variance parameters ($\sigma^2_{w,x}$ and $\sigma^2_{w,y}$).

### Data Analysis

- To assess the performance of each method, the following metrics were calculated:
  - Fraction of experiments in which the estimated coverage rate under each particular approach was the closest to pre-specified confidence rate of 95% among all candidate approaches (i.e., the "best" method)
  - The median coverage rate across all experiments
  - The median discrepancy between the estimated coverage rate across experiments and the pre-specified confidence rate of 95%
- Additionally, the coverage rate of each method was compared under each combination of within- and between-study correlation values for *N* = 20 in a 4 × 4 bar plot.
- For a practical example, the three SE-based 95% CI generation methods were applied to a dataset from a published trial-level surrogacy analysis between hazard ratios (HRs) of event-free survival (EFS) and overall survival (OS) in neoadjuvant treatment of non-small cell lung cancer [Ostoros et al. (2023)].[7] Estimated SE-based 95% CIs were then compared to the authors' bootstrapped 95% CI.

## Results

### Simulation Results

- The performance of the four methods are summarized in **Table 1**:
  - The Hill (1973) approach was the best method (i.e., the method with the observed coverage rate closest to 95%) in 78.1% of experiments, with a median coverage rate of 95.02% and a median discrepancy of 0.80 %-points (i.e., on average the difference between the observed and expected [95%] coverage rates was 0.8 %-points).
  - The Kish (1965) approach was the best method in 17.2% of experiments, with a median coverage rate of 97.01% and a median discrepancy of 2.02 %-points.
  - The naïve approach was the best method in only 4.7% of experiments, with a median coverage rate of 92.38% and a median discrepancy of 2.63 %-points.
  - Bootstrapping never emerged as the best method in any of the experiments, with a median coverage rate of 91.91% and median discrepancy of 3.10 %-points.
- In **Figure 1**, the empirical coverage rates obtained from each of the four methods are compared for each combination of between- and within-study correlation for *N* = 20. Consistent with **Table 1**, the coverage rates obtained from the Hill (1973) approach were consistently the closest to 95%.

### Application to a Published Surrogacy Analysis

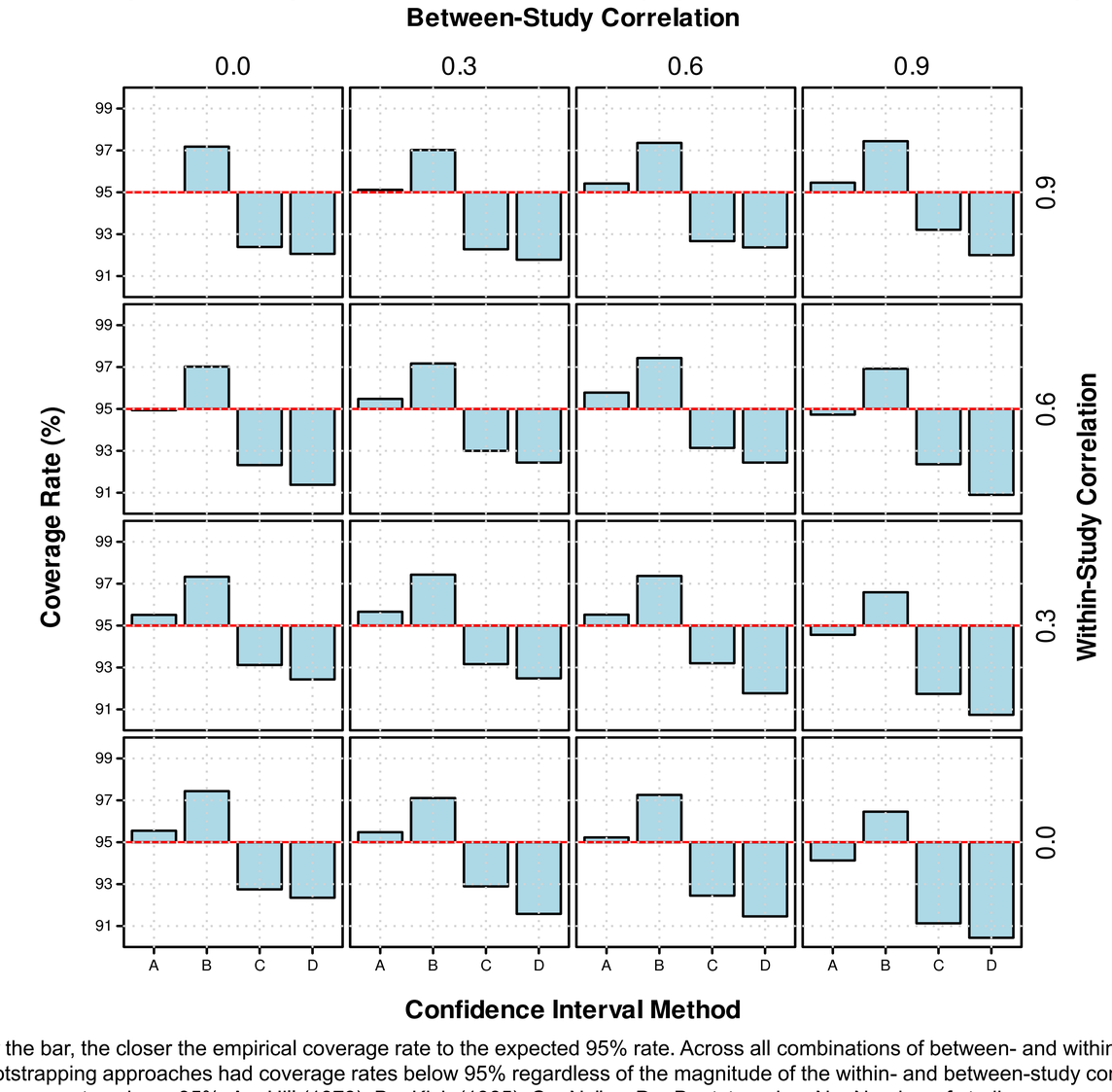
- Ostoros et al. (2023) included eight randomized controlled trials in their trial-level surrogacy analysis of EFS and OS in early-stage non-small cell lung cancer. They estimated a weighted Pearson's correlation of 0.864 between the HRs of EFS and OS, with a 95% CI of 0.809–0.992 estimated from bootstrapping.
- When the three SE-based CI generation methods were applied to the Ostoros et al. (2023) study's evidence base, the Hill (1973) method yielded a 95% CI of 0.268–0.983, the Kish (1965) method yielded a 95% CI of 0.116–0.988, and the naïve method yielded a 95% CI of 0.423–0.976.
- Methods with higher coverage rates in the simulation had broader 95% CIs around the correlation coefficient in the Ostoros et al. (2023) study's evidence base. This observation was consistent with the expectation that 95% CIs are more likely to capture the underlying correlation as they get wider.

**Table 1:** Performance metrics among alternative approaches across simulation experiments.

| | Hill (1973) | Kish (1965) | Naïve | Bootstrapping |
|---|---|---|---|---|
| **Median Coverage (%)[a]** | 95.02 | 97.01 | 92.38 | 91.91 |
| **Median Discrepancy (%-points)[b]** | 0.80 | 2.02 | 2.63 | 3.10 |
| **Best Method (% of experiments)[c]** | 78.1 | 17.2 | 4.7 | 0 |

a. Median coverage rate across experiments; the closer to the expected 95% confidence level, the better the method's performance.
b. Median difference between the estimated coverage rate and the expected 95% confidence level across experiments. Smaller values of discrepancy indicate better performance.
c. Percentage of experiments in which the corresponding method's estimated coverage rate was the closest to pre-specified confidence rate of 95% across all methods. Higher percentage values indicate better performance.

**Figure 1:** Empirical coverage rates among alternative approaches for varying levels of between- and within-study correlations, and *N* = 20.




Note: The shorter the bar, the closer the empirical coverage rate to the expected 95% rate. Across all combinations of between- and within-study correlations, while the naïve and bootstrapping approaches had coverage rates below 95% regardless of the magnitude of the within- and between-study correlations, the Kish approach had coverage rates above 95%. A – Hill (1973), B – Kish (1965), C – Naïve, D – Bootstrapping, N – Number of studies.

## Conclusions

- The results of this simulation study suggest that bootstrapping may produce excessively narrow 95% CIs around the weighted correlation. Using ESS for sample size in the SE calculation may produce more accurate CIs, and has the additional benefit of being easier to implement.
- This finding can assist researchers to produce more reliable results when validating surrogate endpoints at the trial-level.
- Limitations of this work include its sole focus on 95% CIs rather than CIs at varying levels of confidence as well as the assumptions of bivariate normality of treatment effects on the surrogate and true endpoints, equal within-study sampling variances for both endpoints, and a single fixed within-study correlation for all studies.

## References

1. Taylor RS and Elston J. *NIHR Health Technology Assessment*. 2009. 13(8):1–50.
2. Wheaton L and Bujkiewicz S. *Int J Technol Assess Health Care*. 2025. 41(1):e11.
3. Fisher, RA. Statistical Methods for Research Workers. 1925. Oliver and Boyd.
4. Hill MO. *Ecology*. 1973. 52(2):427–432.
5. Kish L. Survey Sampling. 1965. John Wiley & Sons.
6. Sanderson C and Curtin R. 2025 17th ICCAE, 303–307.
7. Ostoros et al. *Expert Rev Anticancer Ther*. 2023. 23(12):1305–1313.

## Acknowledgments

Authors report employment with Evidinno Outcomes Research Inc. (Vancouver, BC, Canada). This research was conducted during Murat Kurt's employment with Evidinno Outcomes Research Inc.

EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

Presented at ISPOR Europe 2025; November 9–12; Glasgow, Scotland, UK.

Email: mfazeli@evidinno.com