abbvie

Evaluating Frequentist and Bayesian Confidence Intervals for Study Size Adjusted (SSA) Risk Difference

Benjamin Duncan

Joint work with Pratyusa Datta (Safety Statistics Intern) and

Cheng Li

Immunology Safety Statistics

November 3, 2025

Agenda / Outline

- Motivation, Problem Statement and Additional Background Information
- Methods
- Simulation Plan to Investigate Methods
- Simulation Results
- Discussion / Conclusions
- References

Statistical Sciences

Motivation
Problem Statement and Background

abbvie

Motivation

Meta-Analyses of Randomized Controlled Clinical Trials to Evaluate the Safety of Human Drugs or Biological Products Guidance for Industry

(*Draft Guidance, U.S. Department of Health and Human Services, FDA, CDER, CBER, November 2018, Drug Safety*)

- If 1:1 randomization is not employed in all trials included in meta-analysis, Simpson's paradox can occur in a raw pooled analysis. The analysis should be stratified by trial (this maintains randomized comparisons of Drug A to Drug B [*or placebo*]), (Section V.B. pg. 14-15)

Many real life examples of Simpson's Paradox exist in the literature:

- 1973 – UC Berkeley Gender Bias Graduate School Admission case[a]
- 1986 – Kidney Stone Treatment (Chiang et al, Br Med Journal)[a]
- 1996 – Appleton, French, Vanderpump. Effects of Smoking [Berman, 2012]

[a] (https://en.wikipedia.org/wiki/Simpson%27s_paradox)

Motivation: Simpson's Paradox Example

- Numerical Example of Simpson's Paradox:

  (from Armitage and Berry [1987], Statistical Methods in Medical Research, p. 382)

Responders

Study	Drug A x / n (%)	Drug B x / n (%)	Difference (A – B)
Study 1	150 / 300 (50.0%)	500 / 1200 (41.7%)	8.3%
Study 2	200 / 1000 (20.0%)	50 / 300 (16.7%)	3.3%
Raw Total	350 / 1300 (26.9%)	550 / 1500 (36.7%)	-9.8%
Adjusted			5.9%

Problem Statement

- The Study Size Adjusted (SSA) Method (Crowe et al, 2016) has now been commonly requested by the FDA as a method for combining studies for assessment of risk. AbbVie has incorporated SSA methods in multiple programs.

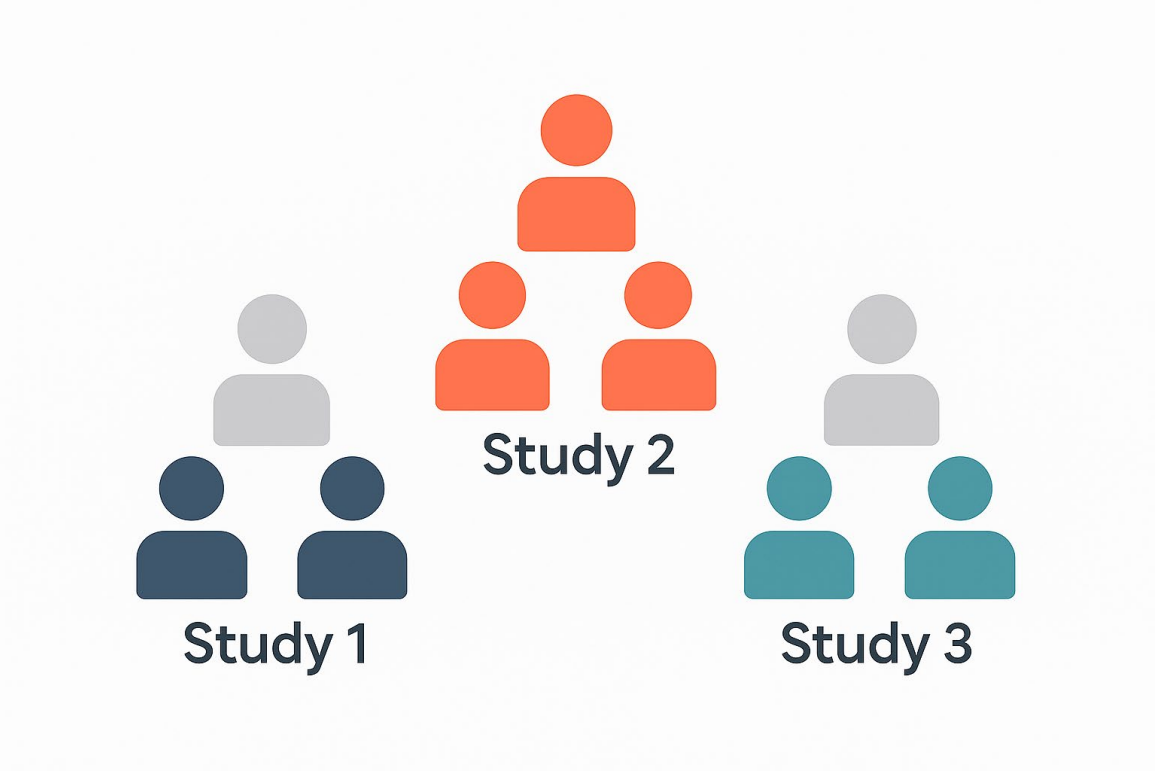
- The computation of the SSA Risk Difference is straight forward and easily programmed. However, there are no established and agreed upon methods for corresponding confidence interval calculations. Therefore, a robust method that can handle large incidence proportions as well as small incidence proportions arising from sparse data (including cases of 0 events) must be formulated. The method, should have adequate coverage (e.g. >= 95%), good Type I error rate (e.g. <= 0.025 [1-sided] and <= 0.05 [2-sided]), and adequate statistical power to detect true differences. In addition, the method formulated for the integrated studies should be harmonious with methods used for individual studies.

Background

Assessing the **risk** of **adverse events (AEs)** associated with drug products requires **combining multiple studies.**

The **Study Size Adjusted (SSA)** method by Crowe et al. (2016) has been commonly requested by the **FDA** to compute **risk difference** combining multiple studies. The weights are based upon arithmetic means. (Note Mantel-Haenszel weights are based upon harmonic means).

There is no agreed upon method to obtain the corresponding **confidence interval (CI).**



Crowe, B. Chuang-Stein, C. Lettis S. Brueckner, A. (2016). Reporting Adverse Drug Reactions in Product Labels. *Therapeutic Innovation & Regulatory Science*; 50(4): 455-463.

Background

Our **objective** is to:

- Formulate a **robust** method to calculate the CI for difference in
  - **Incidence proportion (IP):** Ratio of number of participants who experience an AE to total number of participants.
  - **Incidence Rate (IR)**: Ratio of number of participants who experience an AE to total time at risk

  between treatment and placebo group, combining multiple studies.
- Achieve adequate **coverage**, statistical **power** to detect true differences and low **type 1 error**.
- To overcome the statistical challenges posed by **low event counts** for **rare AE** analysis.

Statistical Sciences

Methods

abbvie

Methods: Outcome and Parameters of Interest

- M studies.
- Data structure of study i for $i = 1, \ldots, M$:

	No. of participants with AE of interest	Total no. of participants	Total time at risk	Incidence Proportion (IP)	Incidence Rate (IR)
Treatment	y_{i1}	n_{i1}	T_{i1}	y_{i1}/n_{i1}	y_{i1}/T_{i1}
Placebo	y_{i0}	n_{i0}	T_{i0}	y_{i0}/n_{i0}	y_{i0}/T_{i0}

- Parameters of interest:
  - Difference in IP between treatment and placebo.
  - Difference in IR between treatment and placebo.

Methods: SSA Weighting (example)

Study	PBO (N)	TRT (N)	Total (N)	SSA Weight
S1	50	50	100	100 / 1,000 = 0.10
S2	100	200	300	300 / 1,000 = 0.30
S3	150	450	600	600 / 1,000 = 0.60
Totals			1,000	

- Arithmetic means:
  - S1: (50 + 50) / 2 = 50; S2: (100 + 200) / 2 = 150; S3: (150 + 450) / 2 = 300
  - Total = 50 + 150 + 300 = 500
  - W1 = 50 / 500 = 0.10; W2 = 150 / 500 = 0.30; W3 = 300 / 500 = 0.60

Statistical Sciences

Methods
(CI for differences in IP)

abbvie

Methods Considered (Overview) – Incidence Proportion

- (1) Traditional Mantel-Haenszel Weighted Method with and without continuity correction (provided as a reference only)
- (2) Asymptotic (Wald) SSA Weighting Method with and without continuity correction
- (3) Klingenberg Method (Mantel-Haenszel weights) (provided as a reference only)
- (4) Modified Klingenberg Method (SSA weights)
- (5) Hybrid Quasi-Exact / Asymptotic (SSA weights)
- (6) Modified Stratified Wald Test (SSA weights)
- (7) Modified Stratified T-test (SSA weights)
- (8) Bayesian without Heterogeneity
- (9) Bayesian with Heterogeneity

Methods Considered (Overview) – Incidence Proportion

- (1) Traditional Mantel-Haenszel Weighted Method ~~with and~~ without continuity correction (provided as a reference only)
- (2) Asymptotic (Wald) SSA Weighting Method ~~with and~~ without continuity correction
- ~~(3) Klingenberg Method (Mantel-Haenszel weights) (also provided as a reference)~~
- ~~(4) Modified Klingenberg Method (SSA weights)~~
- (5) Hybrid Quasi-Exact / Asymptotic (SSA weights)
- (6) Modified Stratified Wald Test (SSA weights)
- (7) Modified Stratified T-test (SSA weights)
- (8) Bayesian without Heterogeneity
- (9) Bayesian with Heterogeneity

Methods Investigated and Displayed (Overview) – Incidence Proportion

- (1) Traditional Mantel-Haenszel Weighted Method without continuity correction (provided as a reference only)
- (2) Asymptotic (Wald) SSA Weighting without continuity correction
- (3) Hybrid Quasi-Exact / Asymptotic (SSA weights)
- (4) Modified Stratified Wald Test (SSA weights)
- (5) Modified Stratified T-test (SSA weights)
- (6) Bayesian without Heterogeneity
- (7) Bayesian with Heterogeneity

Methods: CI for Difference in IP

1. Mantel-Haenszel Weights (provided as a reference only):

- Traditional approach for clinical trials.
  - ❖ Weighting as defined by Mantel and Haenszel (1959) with variance estimator applied/modified by Sato (1989) – method used in SAS Proc Freq when invoking the cl=MH option in the commonriskdiff statement option of the Tables statement.
  - ❖ No CI is produced if both arms have zero events in all studies.

$100(1-\alpha)\%$ CI for the difference in IP:

$\sum_{i=1}^{M} w_i \hat{d}_i \pm z_{1-\frac{\alpha}{2}}\, sd(\sum_{i=1}^{M} w_i \hat{d}_i)$, where $\hat{d}_i$ is the difference in IP for study i,

and $sd\begin{pmatrix}\sum_{i=1}^{M} w_i \\ \hat{d}_i\end{pmatrix} = \sqrt{\sum_{i=1}^{M} w_i^2\, Var(\hat{d}_i)}$ and $w_i = \frac{\left(\frac{1}{n_{i1}} + \frac{1}{n_{i0}}\right)^{-1}}{\sum_{i=1}^{M}\left(\frac{1}{n_{i1}} + \frac{1}{n_{i0}}\right)^{-1}}$.

Mantel, N. and Haenszel, W. (1959). Statistical Aspects of Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*; 22:719-748.
Sato, T. (1989). On the Variance Estimator of the Mantel-Haenszel Risk Difference. *Biometrics*; 45: 1323-1324. Letter to the editor.

Methods: CI for Difference in IP

2. Asymptotic Wald Study Size Adjusted (SSA) Weights:

- SSA weights per Crowe et al. (2016)
  - The classic method for computing variance for a linear combination for a discrete distribution (i.e. π_1 - π_2), with SSA weighting, is used to compute a standard error. The normal approximation along with the standard error estimate is used compute the 95% CI.
  - Works well for large sample sizes and moderate to large IP.
  - No CI is produced if both arms have zero events in all studies.

$100(1-\alpha)\%$ CI for the difference in IP:

$$\sum_{i=1}^{M} w_i \hat{d}_i \pm z_{1-\frac{\alpha}{2}}\, sd\left(\sum_{i=1}^{M} w_i \hat{d}_i\right), \text{ where } \hat{d}_i = \frac{y_{i1}}{n_{i1}} - \frac{y_{i0}}{n_{i0}} \text{ and}$$

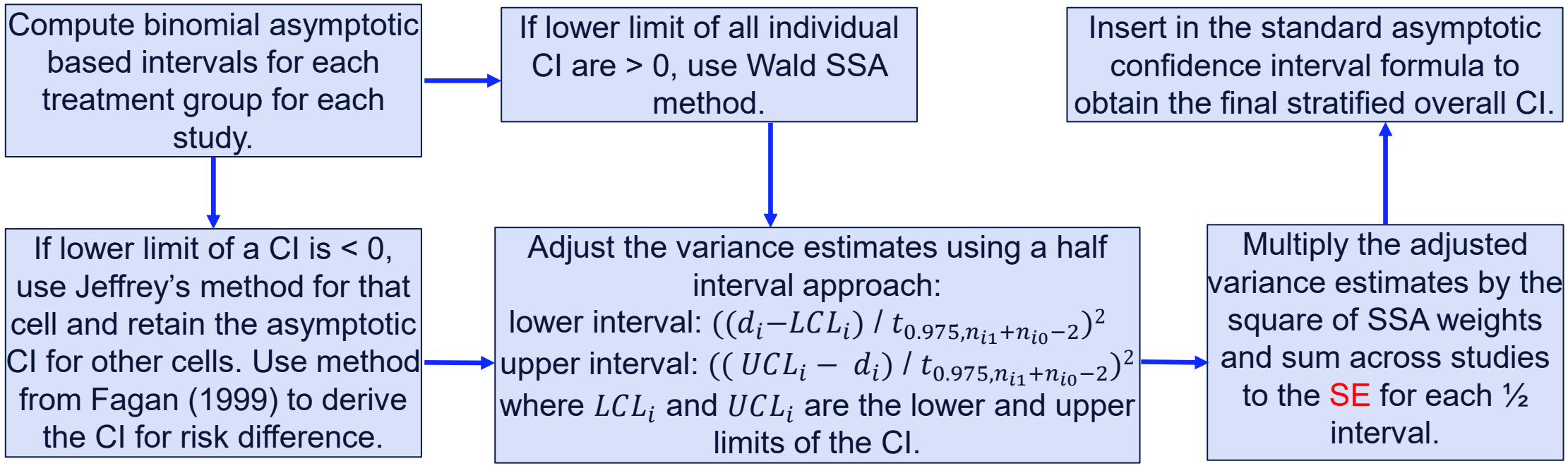
$$sd\left(\sum_{i=1}^{M} w_i\ \hat{d}_i\right) = \sqrt{\sum_{i=1}^{M} w_i^2\ Var(\hat{d}_i)} \text{ and } w_i = \frac{n_{i1}+n_{i0}}{\sum_{i=1}^{M}(n_{i1}+n_{i0})}.$$

Crowe, B. Chuang-Stein, C. Lettis S. Brueckner, A. (2016). Reporting Adverse Drug Reactions in Product Labels. *Therapeutic Innovation & Regulatory Science*; 50(4): 455-463.

Methods: CI for Difference in IP

3. Hybrid Quasi-Exact / Asymptotic SSA Weights:

- Incorporates both exact and asymptotic methods conditional on rates and sample size.
- Likely requires vetting/approval from industry and regulatory agencies.



Methods: CI for Difference in IP

4. **Modified Stratified Wald (with +2/+4 adjustment):**

- A modification of Wald's method with +2/+4 adjustments to the numerator and denominator (Sui et al., 2021) with SSA weights substituted for the MH weights.

- Higher than nominal Type I error when sparsity is an issue.

$100(1-\alpha)\%$ CI for the difference in IP:

$\sum_{i=1}^{M} w_i \hat{d}_i \pm z_{1-\frac{\alpha}{2}}\, sd(\sum_{i=1}^{M} w_i \hat{d}_i)$, where $\widetilde{d}_i = \frac{y_{i1}+w_i}{n_{i1}+2w_i} - \frac{y_{i0}+w_i}{n_{i0}+2w_i}$ and

$sd\left(\sum_{i=1}^{M} w_i\ \widetilde{d}_i\right) = \sqrt{\sum_{i=1}^{M} w_i^2\ Var(\widetilde{d}_i)}$ and $w_i = \frac{n_{i1}+n_{i0}}{\sum_{i=1}^{M}(n_{i1}+n_{i0})}$.

Sui J, Jiao J, Sun Y, Liu J, Bastero R, Koch G. (2021). Evaluation of alternative confidence intervals to address non-inferiority through the stratified difference between proportions. Pharm Stat. 2021; 20(1): 146-162.

Methods: CI for Difference in IP

5. **Modified Stratified t-distribution with +2/+4 adjustment:**

- Variation of the Modified Stratified Wald method with SSA weights, uses t distribution quantiles.

- Sui et al. (2021) demonstrate that this method controls nominal type 1 error for sparse data.

$100(1-\alpha)\%$ CI for the difference in IP:

$$\sum_{i=1}^{M} w_i \hat{d}_i \pm t_{1-\frac{\alpha}{2}}\, sd\left(\sum_{i=1}^{M} w_i \hat{d}_i\right), \text{ where } \widetilde{d}_i = \frac{y_{i1}+w_i}{n_{i1}+2\,w_i} - \frac{y_{i0}+w_i}{n_{i0}+2\,w_i} \text{ and}$$

$$sd\left(\sum_{i=1}^{M} w_i\ \widetilde{d}_i\right) = \sqrt{\sum_{i=1}^{M} w_i^2\ Var(\widetilde{d}_i)} \text{ and } w_i = \frac{n_{i1}+n_{i0}}{\sum_{i=1}^{M}(n_{i1}+n_{i0})}.$$

(note: see Sui et al., 2021 for calculation of v)

Sui J, Jiao J, Sun Y, Liu J, Bastero R, Koch G. (2021). Evaluation of alternative confidence intervals to address non-inferiority through the stratified difference between proportions. Pharm Stat. 2021; 20(1): 146-162.

Methods: CI for Difference in IP

6. Bayesian Beta without Heterogeneity:

- Suitable for sparse data arising from low IP according to Hong et al. (2021).
- Assumes homogeneity across studies.

Model:

For $i = 1, \ldots, M$,

$$y_{i1} \sim \text{Binomial}(n_{i1}, p_{i1}) \text{ and } y_{i0} \sim \text{Binomial}(n_{io}, p_{i0})$$

$$p_{i1} \sim \text{Beta}(1,1) \text{ and } p_{i0} \sim \text{Beta}(1,1)$$

Hong H, Wang C, Rosner GL. (2021). Meta-analysis of rare adverse events in randomized clinical trials: Bayesian and frequentist methods. Clin Trials;18(1):3-16.

Methods: CI for Difference in IP

7. Bayesian Beta with Heterogeneity:

– Suitable for sparse data arising from low IP according to Hong et al. (2021).

Model:

For $i = 1, \ldots, M$,

$$y_{i1} \sim \text{Binomial}(n_{i1}, p_{i1}) \text{ and } y_{i0} \sim \text{Binomial}(n_{io}, p_{i0})$$

$$p_{ik} \sim \text{Beta}(U_k V_k, (1 - U_k) V_k), \text{for } k = 0, 1,$$

$$V_k \sim \text{Inverse Gamma}(1, 0.01), \text{for } k = 0, 1,$$

$$U_k \sim \text{Beta}(1, 1), \text{for } k = 0, 1.$$

Statistical Sciences

Methods
(CI for differences in IR)

abbvie

Methods: CI for Difference in IR

1. Asymptotic SSA Weighting Method:

- Routinely used to compute CI in clinical trials.
- Works well for large sample sizes and moderate to large IR.
- No CI is produced if both arms have zero events in all studies.

$100(1-\alpha)\%$ CI for the difference in IR:

$$\sum_{i=1}^{M} w_i \ (\widehat{\lambda_{i1}} - \widehat{\lambda_{i0}}) \pm z_{1-\frac{\alpha}{2}} \sqrt{\sum_{i=1}^{M} w_i^2 \ (V(\widehat{\lambda_{i1}}) + V(\widehat{\lambda_{i0}}))},$$

where $\hat{\lambda}_{ij} = \frac{y_{ij}}{T_{ij}}$, $V(\widehat{\lambda_{i1}}) = \frac{y_{ij}}{T_{ij}^2}$ and $w_i = \frac{\sum_j T_{ij}}{\sum_i \sum_j T_{ij}}$ for $j = 0, 1$ and $i = 1, \ldots, M$.

Methods: CI for Difference in IR

2. **Bayesian Right Censored Poisson with Non-informative Prior**:

For the jth patient in the ith study, where $i = 1, \dots, M$ and $j = 1, \dots, n_i$,

- $X_{ij} = 1$ if the patient experiences an AE and 0 otherwise.
- t_{ij} is the time at risk.
- $\delta_{ij} = 1$ for treatment and 0 for placebo.

Model:

$$X_{ij}|t_{ij} \sim \text{Bernoulli}(1 - \exp(-(\eta + \Upsilon\delta_{ij})t_{ij})),$$
$$\log \eta \sim N(0, \sigma^2 = 100)$$
$$\log \Upsilon \sim N(0, \sigma^2 = 100)$$

Analysis of sparse AE data might be challenging.

Methods: CI for Difference in IR

3. **Bayesian Right Censored Poisson with Weakly Informative Prior**:

- Puts weak restrictions on the size of the risk difference.

- Suitable for low as well as high IR.

Model:

$$X_{ij}|t_{ij} \sim \text{Bernoulli}(1 - \exp(-(\eta + \Upsilon\,\delta_{ij})t_{ij})),$$
$$\log \eta \sim N(0, \sigma^2 = 50)$$
$$\log \Upsilon \sim N(0, \sigma^2 = 5)$$

Methods: CI for Difference in IR

4. **Bayesian Exponential with Non-informative Prior**:

- For study i, we denote time to the first AE for the jth patient with V_{ij}.
- Right censoring at the end of the follow-up period U_i for study i.
- We observe $Z_{ij} = \min(V_{ij}, U_i)$

Model:
$V_{ij} \sim \text{Exponential}(\eta + \Upsilon\, \delta_{ij})$ and $Z_{ij} = \min(V_{ij}, U_i)$ $\log \eta \sim N(0, \sigma^2 = 100)$ $\log \Upsilon \sim N(0, \sigma^2 = 100)$

Analysis of sparse AE data might be challenging.

Methods: CI for Difference in IR

5. **Bayesian Exponential with Weakly Informative Prior**:

- Puts weak restrictions on the size of the risk difference.

- Suitable for low as well as high IR.

Model:

$$V_{ij} \sim \text{Exponential}\left(\eta + \Upsilon\, \delta_{ij}\right) \text{ and } Z_{ij} = \min(V_{ij}, U_i)$$
$$\log \eta \sim N(0, \sigma^2 = 50)$$
$$\log \Upsilon \sim N(0, \sigma^2 = 5)$$

Statistical Sciences

Simulation Plan to Investigate Methods

abbvie

Simulation Plan to Investigate Methods

- Various scenarios were simulated to assess the coverage probability, Type I error rates, power, and interval length of the various methods.

- The number of adverse events were simulated using the binomial distribution. Both homogenous and heterogenous conditions were assessed. Heterogeneity was introduced by changing the follow-up time, in which response rates were inferred using the exponential distribution (which assumes a constant hazard rate).

- For each simulation run, three trials with two treatment groups (placebo, active treatment) were created. Various sample sizes per trial and various randomization ratios were evaluated.

Simulation Study: Design

- 3 2-arm trials.
- Follow up times of 12 - 48 weeks for each group, trial.
- For IP, number of participants who experience an AE follow binomial.
- For IR, time to first AE follows exponential.
- 10,000 - 50,000 runs (frequentist models)
  500 – 5,000 runs (Bayesian models).
- We consider various simulation scenarios:
  - Balanced and unbalanced randomization ratios (RR), from 1:1 to 3:1.
  - Varying sample sizes.
  - Different IP and IR (small to medium/large)
  - Different risk differences (0 – 1.75 for IR, -0.25% - 4.5% for IP)

Incidence Percentage

Scenario	Rates	RD	RR	Total N
2.1	Homogenous	0	All 1:1	900
1.1	Homogenous	0	Varying	900
3.1	Homogenous	Act >= PBO	Varying	900
4.1	Homogenous	0	Varying	1225
5.1	Homogenous	Act >= PBO	Varying	630

Incidence Rate

Scenario	Rates	RD	RR	Total N
3.7	Homogenous	0	Varying	900
3.8	Homogenous	Act >= PBO	Varying	900
…				

Statistical Sciences

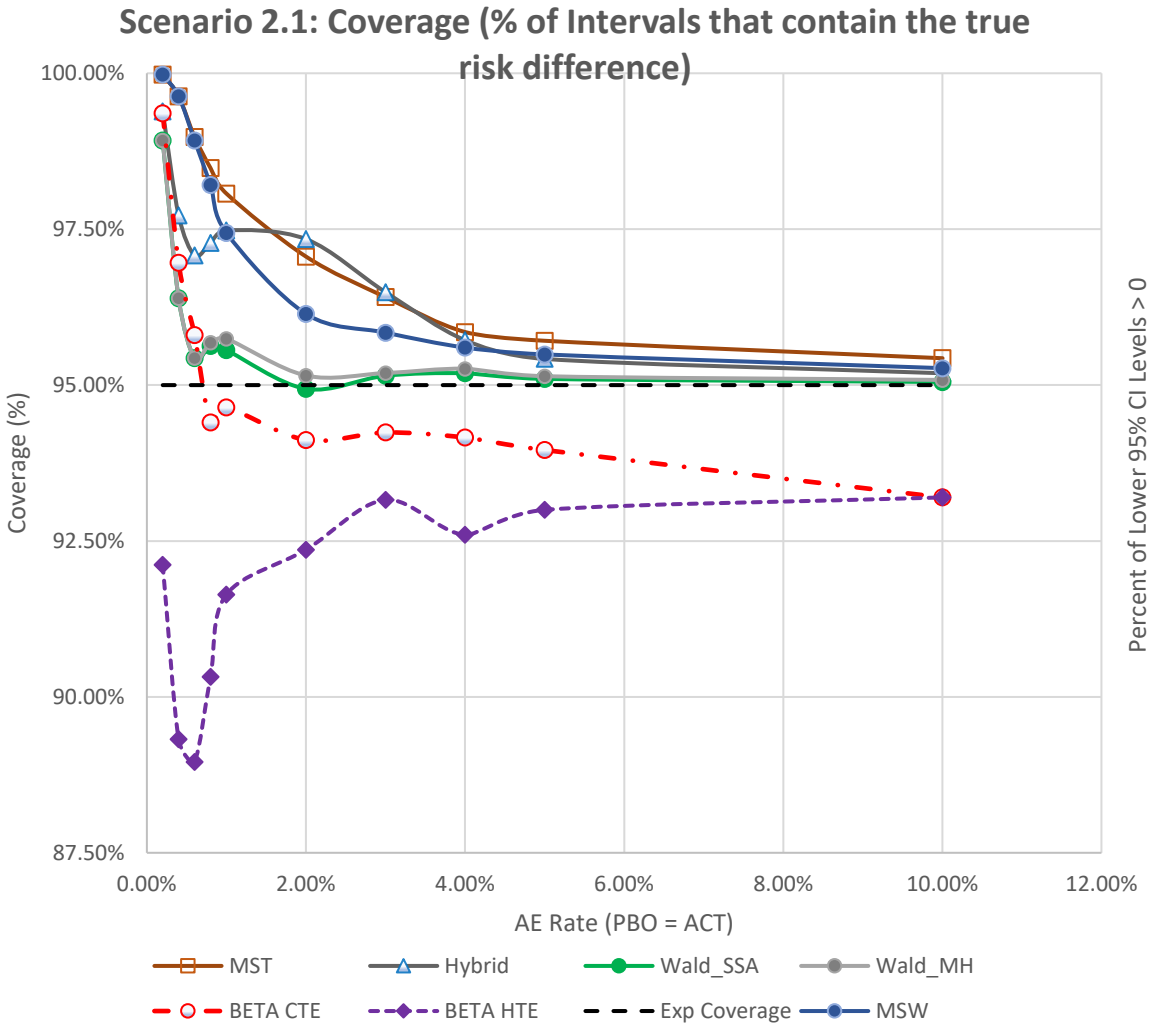
Simulation Results

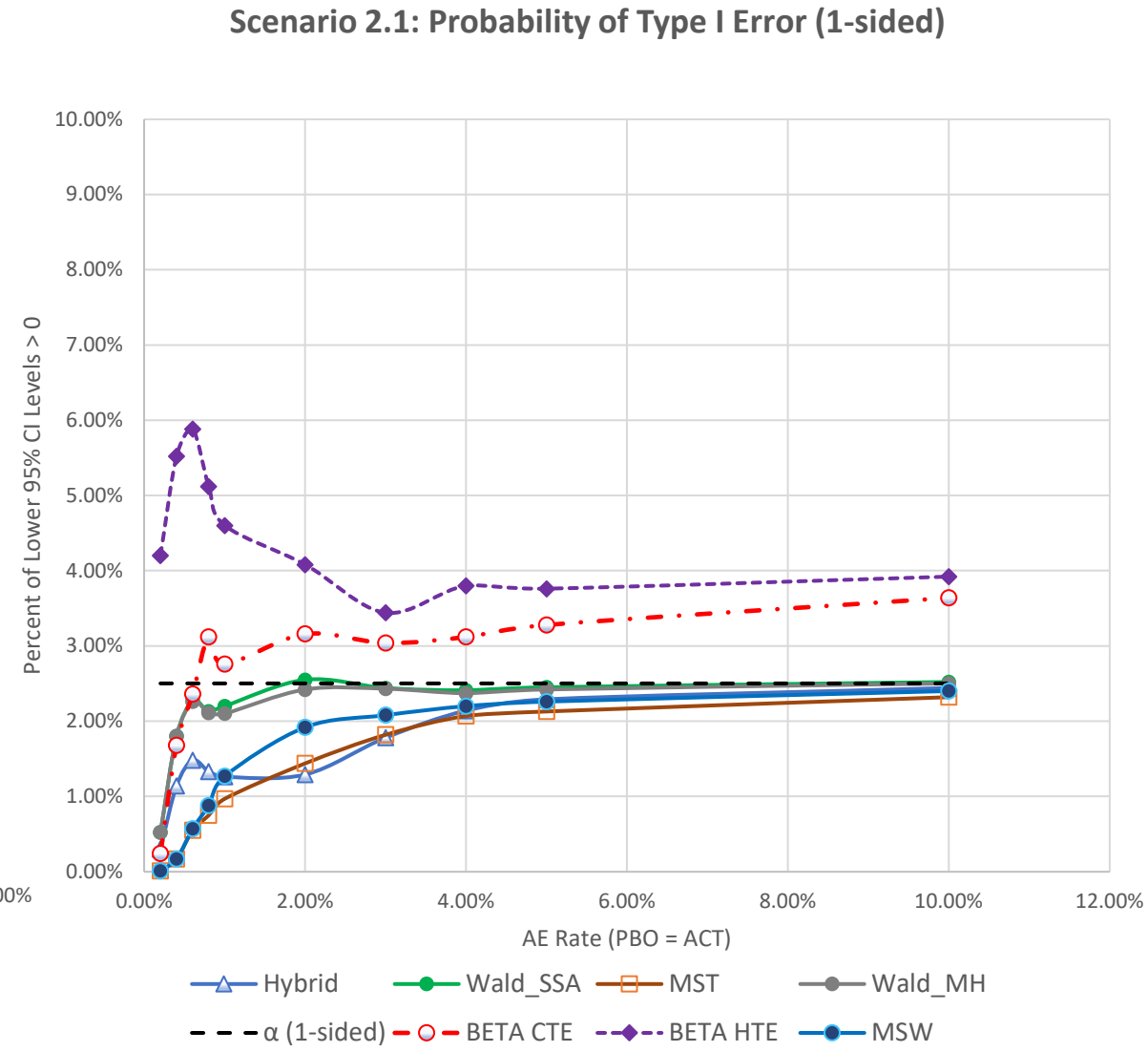
abbvie

Simulation Results: Table to Link Methods to Legend (in subsequent Plots)

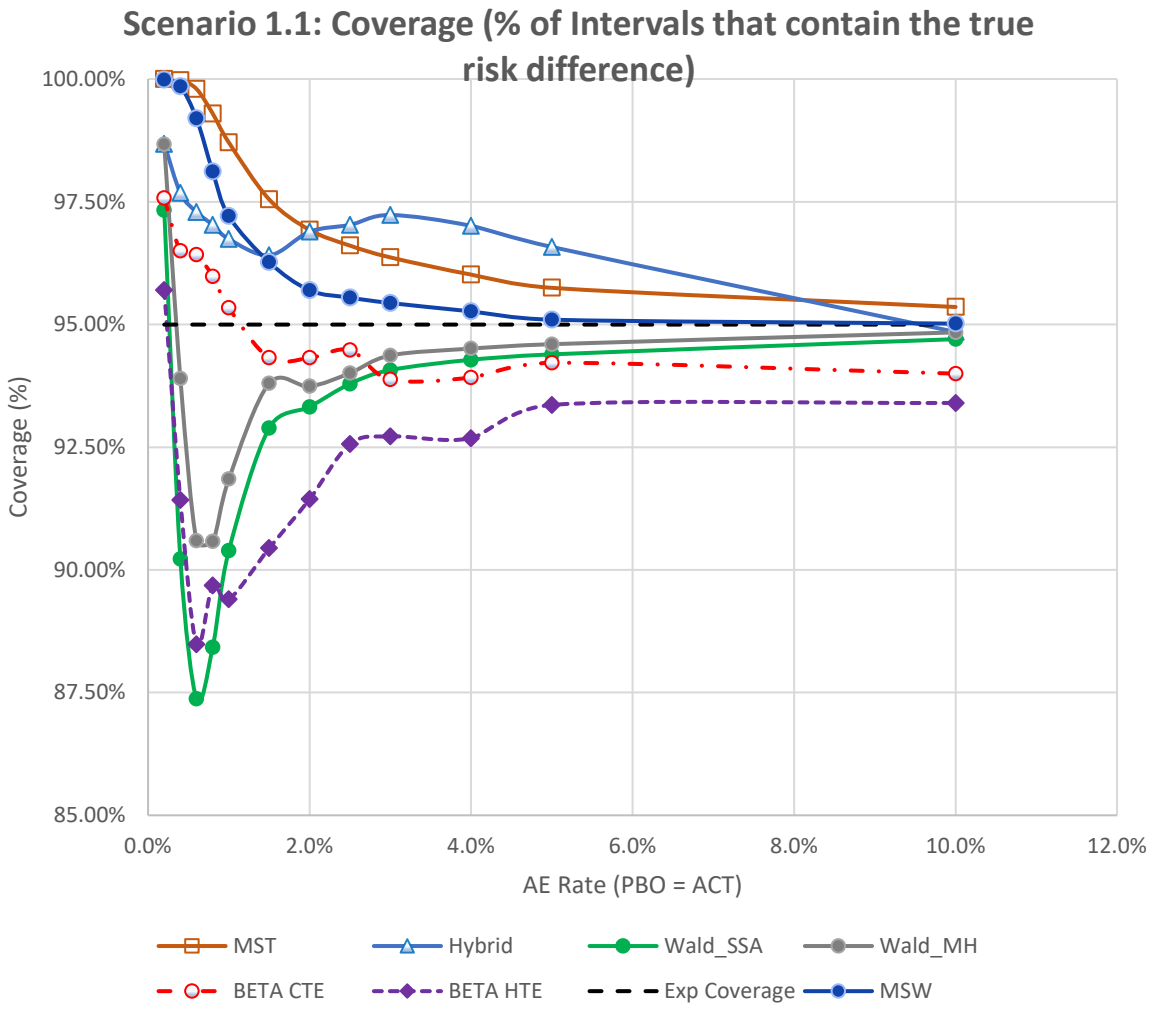
IP Methods	Abbreviation Used in Simulation Plots
(1) Asymptotic (Wald) Mantel-Haenszel Weighted Method	Wald_MH
(2) Asymptotic (Wald) SSA Weighting Method	Wald_SSA
(3) Hybrid Quasi-Exact / Asymptotic (SSA weights)	Hybrid
(4) Modified Stratified Wald (SSA weights)	MSW
(5) Modified Stratified t-distribution (SSA weights)	MST
(6) Bayesian without heterogeneity	BETA_CTE
(7) Bayesian with heterogeneity (SSA weights)	BETA_HTE

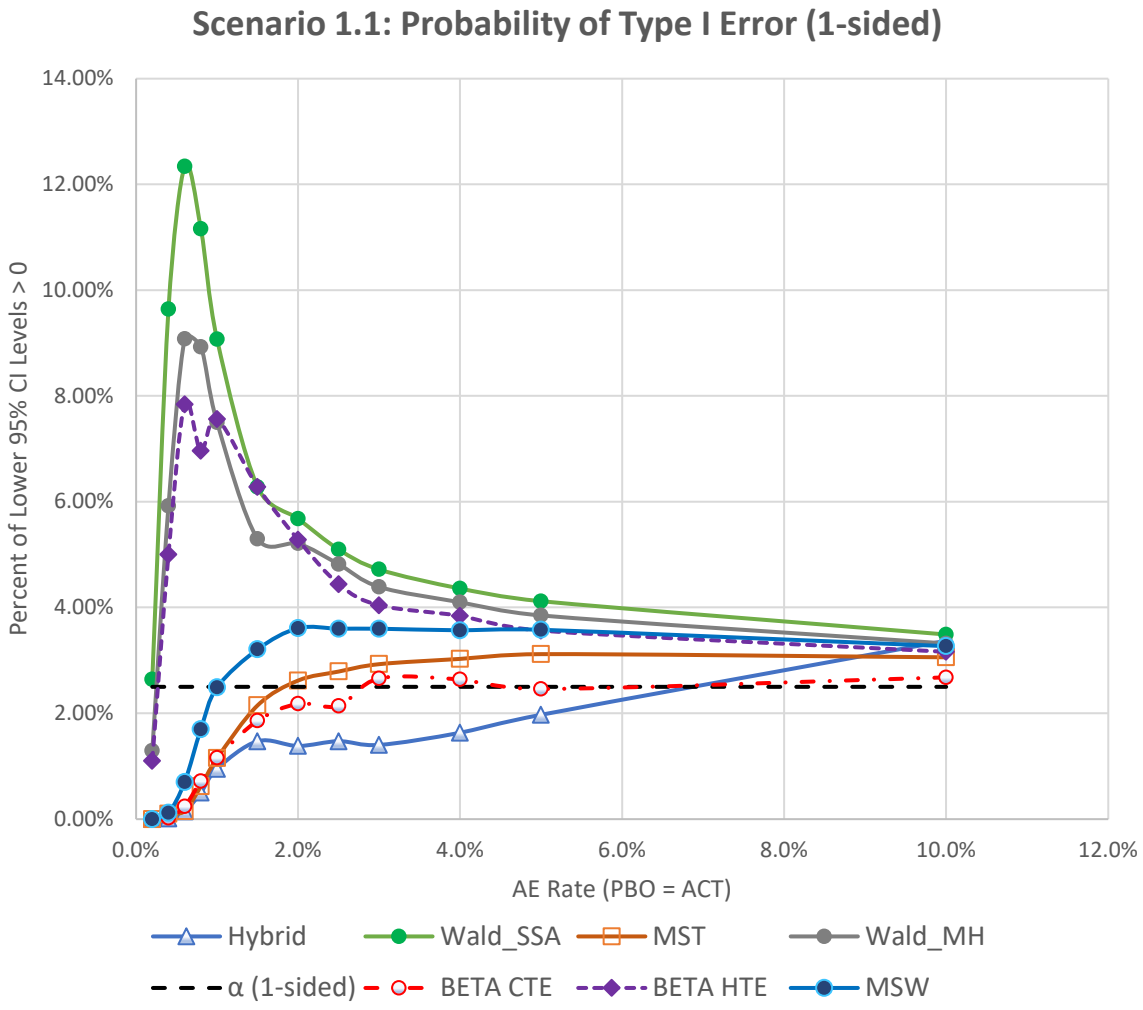
Simulation Results: Risk Difference = 0 (IP from 0.2% to 10.0%)
[Scenario 2.1: Homogenous Rates Across Studies, all 1:1 RRs; N_1.=100 (1:1), N_2. and N_3.= 400 (1:1)]



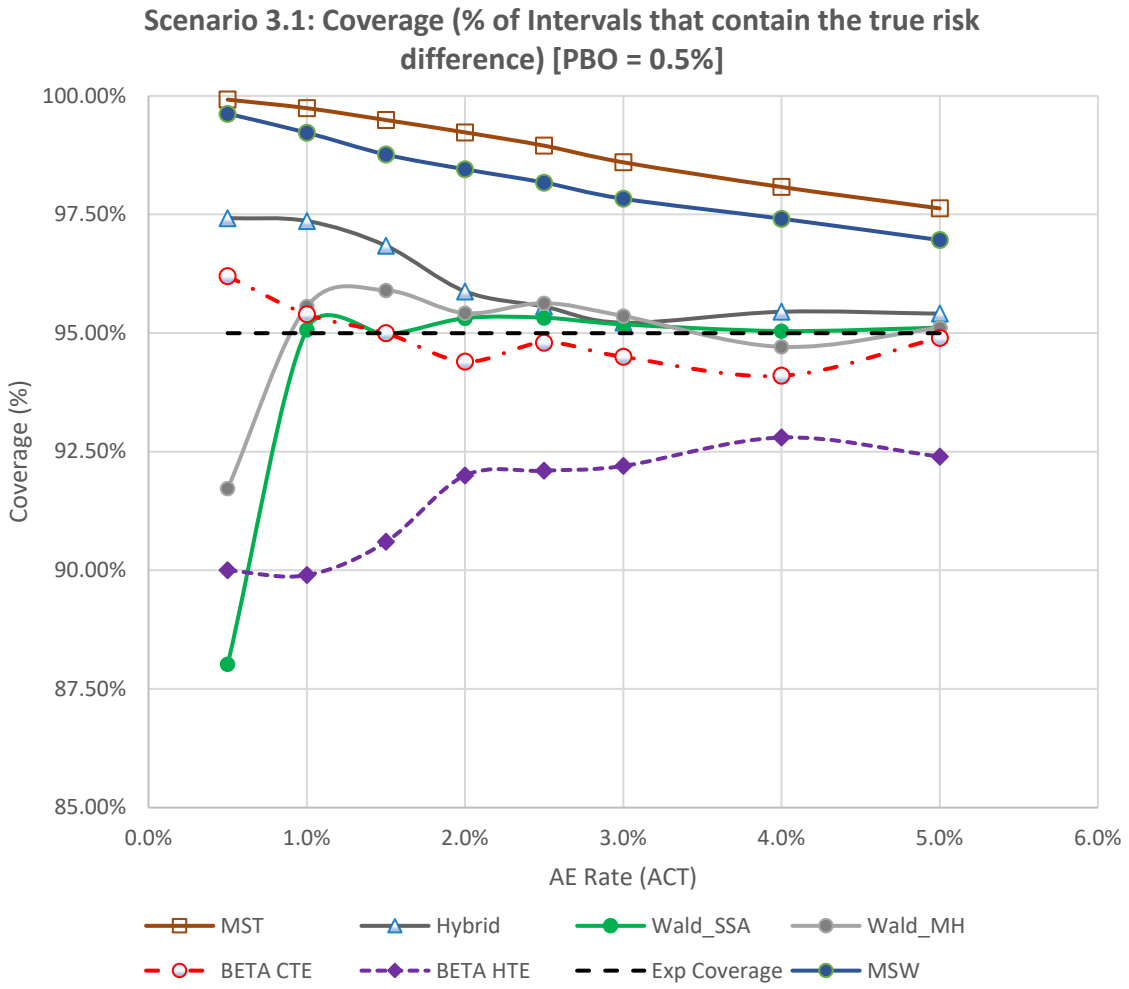


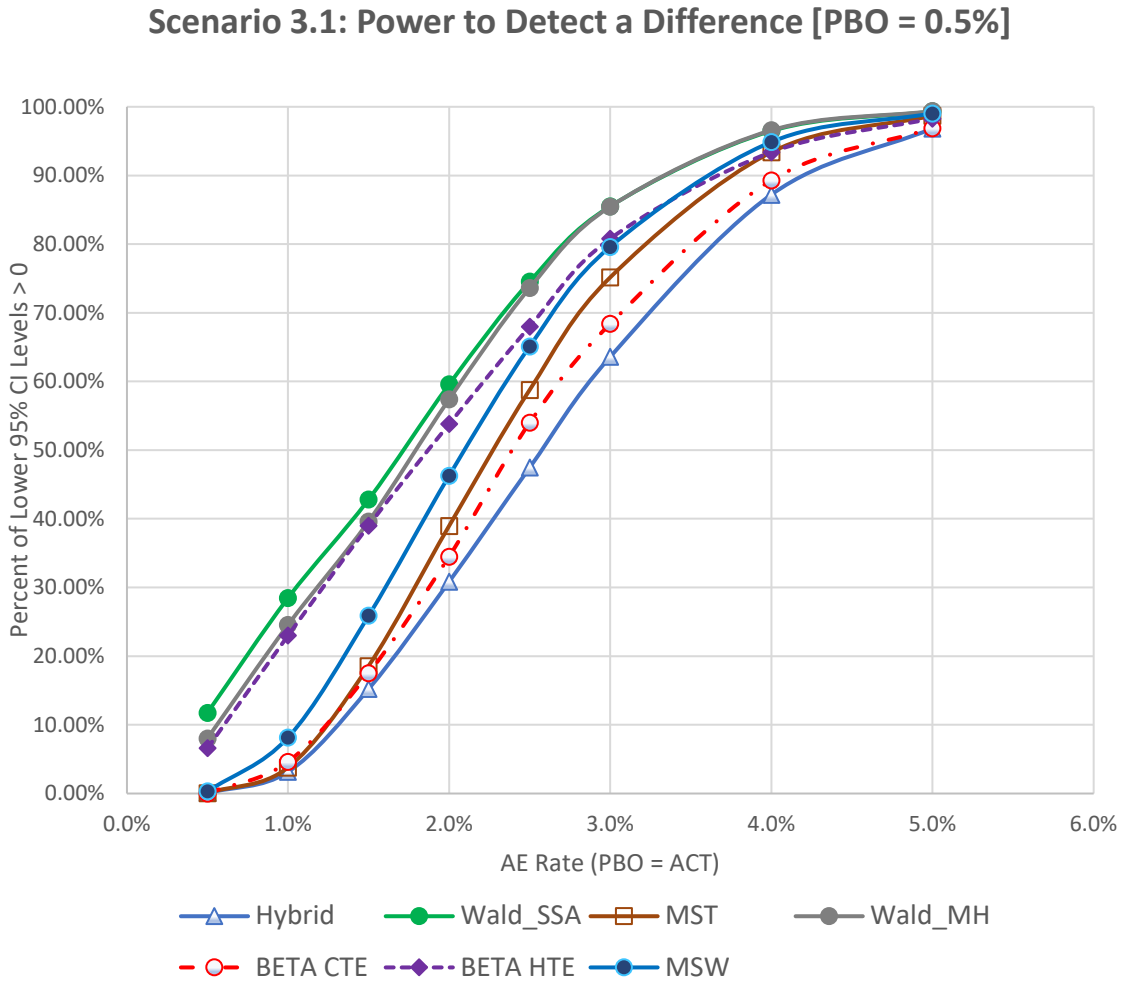
Simulation Results: Risk Difference = 0 (IP from 0.2% to 10.0%)
[Scenario 1.1: Homogenous Rates Across Studies, varying RRs; N_1.=100 (1:1), N_2. and N_3.= 400 (3:1)]



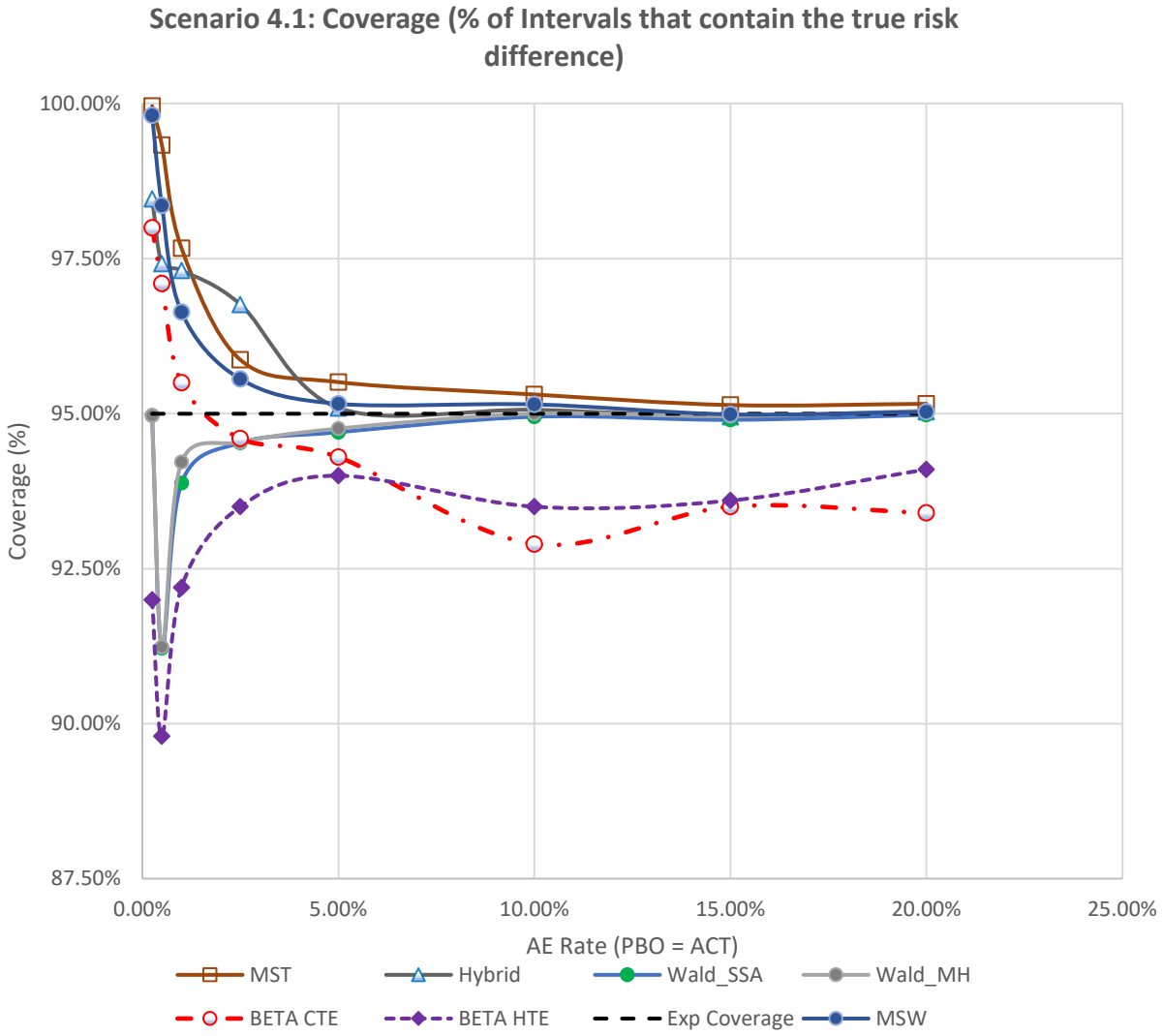


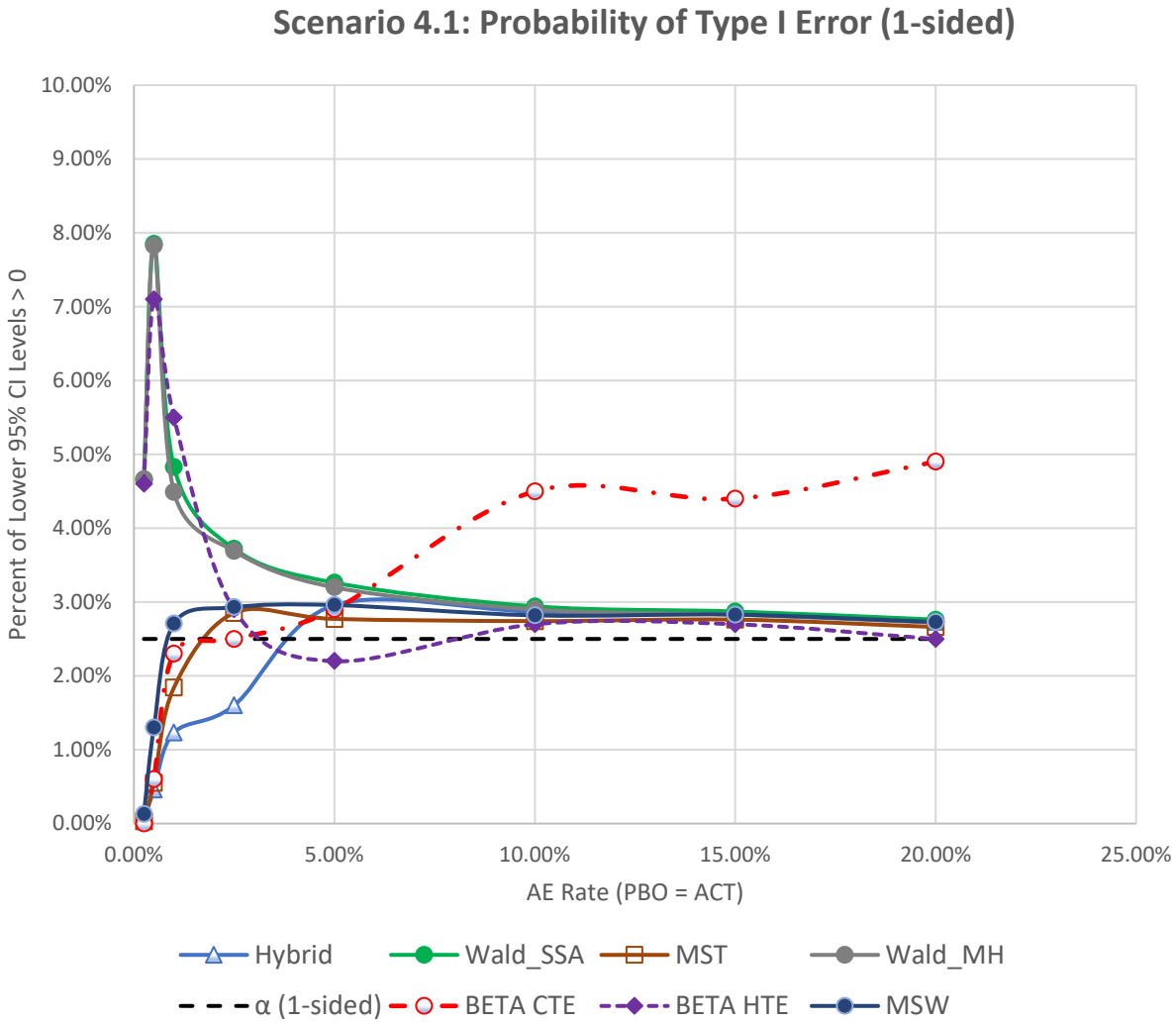
Simulation Results: Risk Difference = 0 to 4.5% (PBO IP = 0.5%)
[Scenario 3.1: Homogenous Rates Across Studies, varying RRs; N_1.=100 (1:1), N_2. and N_3.= 400 (3:1)]



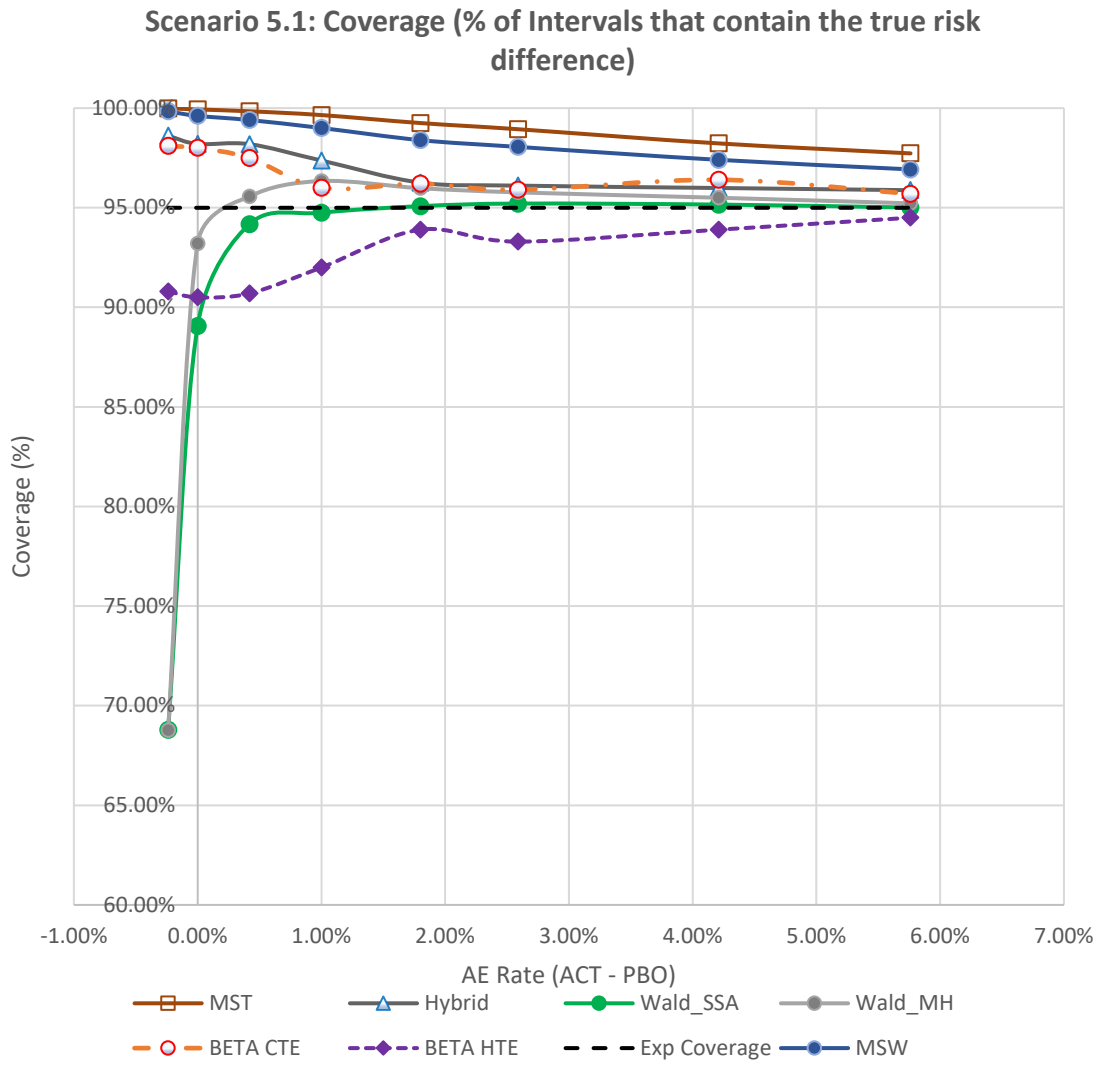


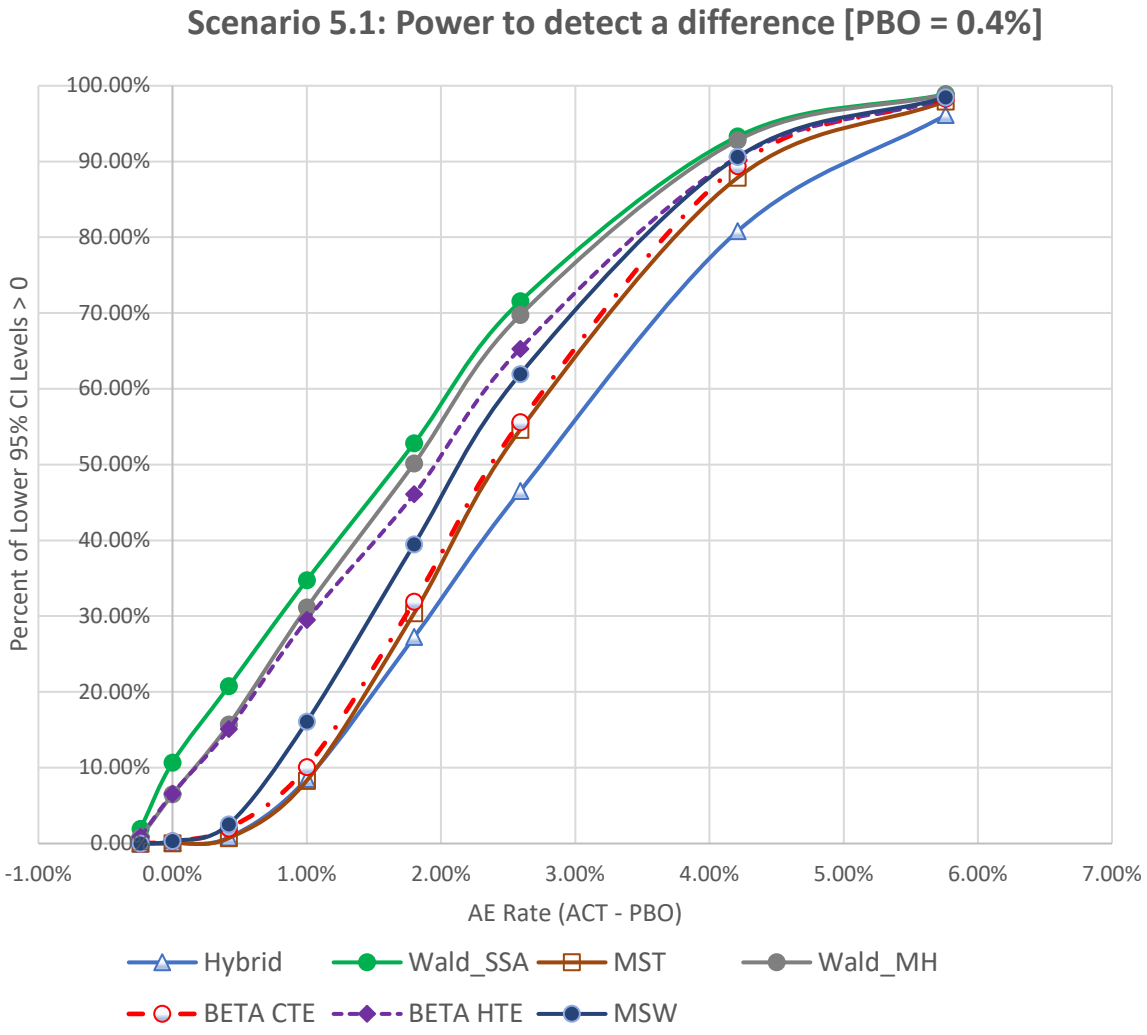
Simulation Results: Risk Difference = 0 (IP from 0.2% to 20.0%)
[Scenario 4.1: Heterogenous Rates Across Studies, varying RRs; N_1.=100 (1:1), N_2. = 600 (2:1), N_3.= 525 (2:1)]





Simulation Results: Risk Difference = -0.25% to 5.75% (PBO IP = 0.4%)
[Scenario 5.1: Heterogenous Rates Across Studies, varying RRs; N_1.=80 (1:1), N_2. = 150 (2:1), N_3.= 400 (3:1)]





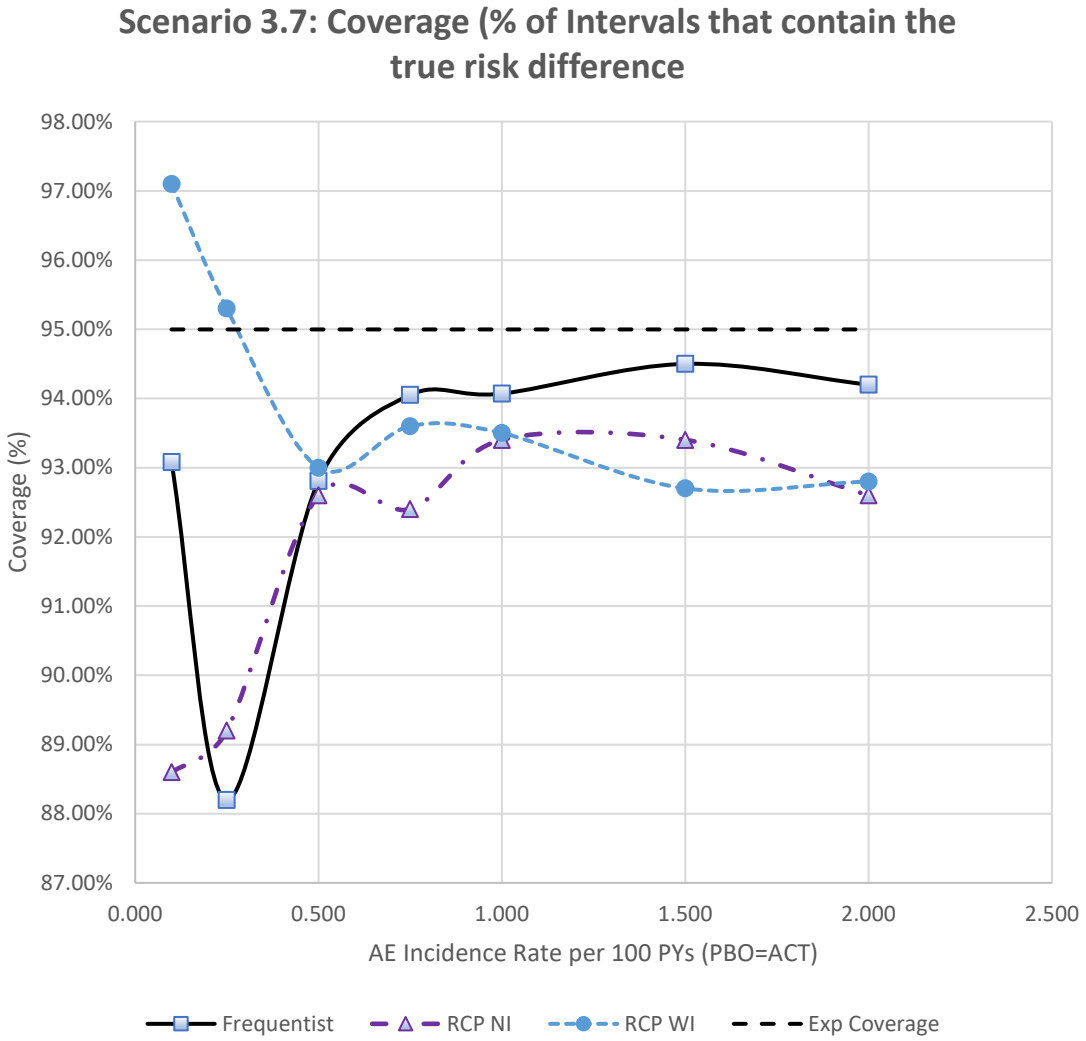
Simulation Study: Summary for difference in IP

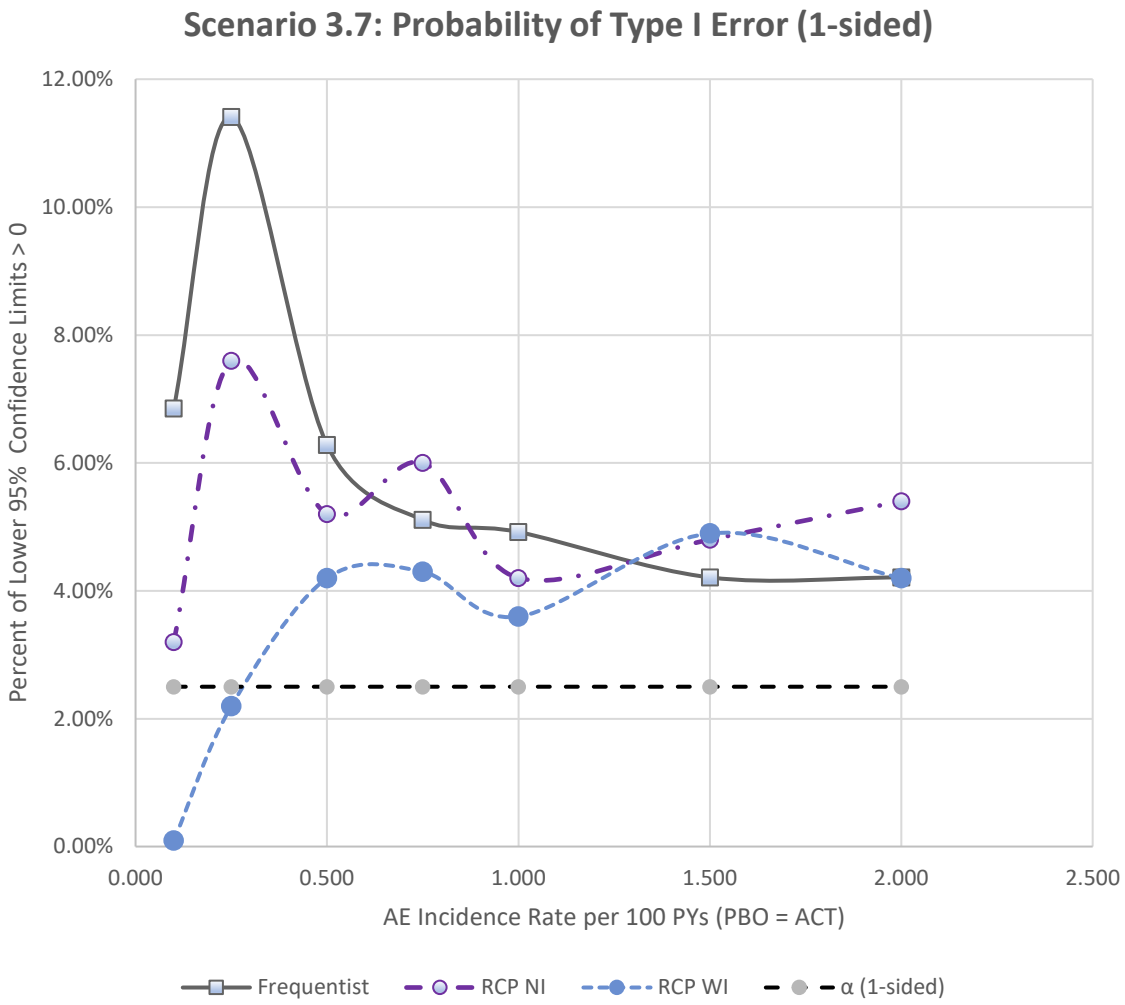
Method	Key Takeaway
Wald_MH	Poor coverage and high type 1 error for small IP for varying (unbalanced) RR.
Wald_SSA	
MSW	**Performs well in all cases**. MSW slightly better power. MST slightly better Type I error control. Hybrid lower power compared to MSW and MST
MST	
Hybrid	
Beta CTE	Mixed results, performs well for homogenous rates, not so well for heterogenous rates.
Beta HTE	Performs poorly in all examined scenarios.

Simulation Results: Table to Link Methods to Legend (in subsequent Plots)

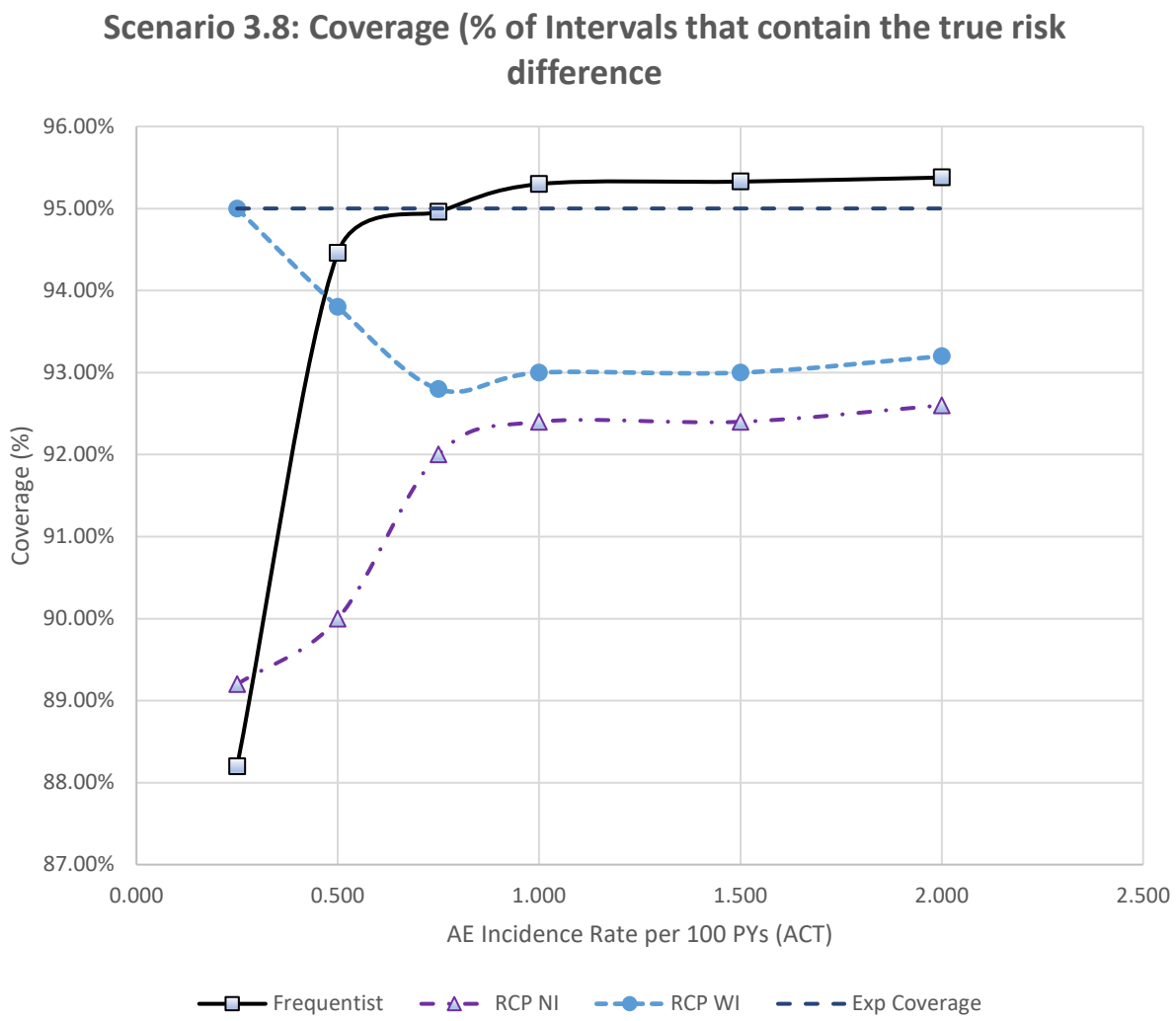
IR Methods	Abbreviation Used in Simulation Plots
(1) Asymptotic SSA Weighting Method	Frequentist
(2) Bayesian Right Censored Poisson with Non-informative Prior	RCP NI
(3) Bayesian Right Censored Poisson with Weakly Informative Prior	RCP WI
(4) Bayesian Exponential with Non-informative Prior	
(5) Bayesian Exponential with Weakly Informative Prior	

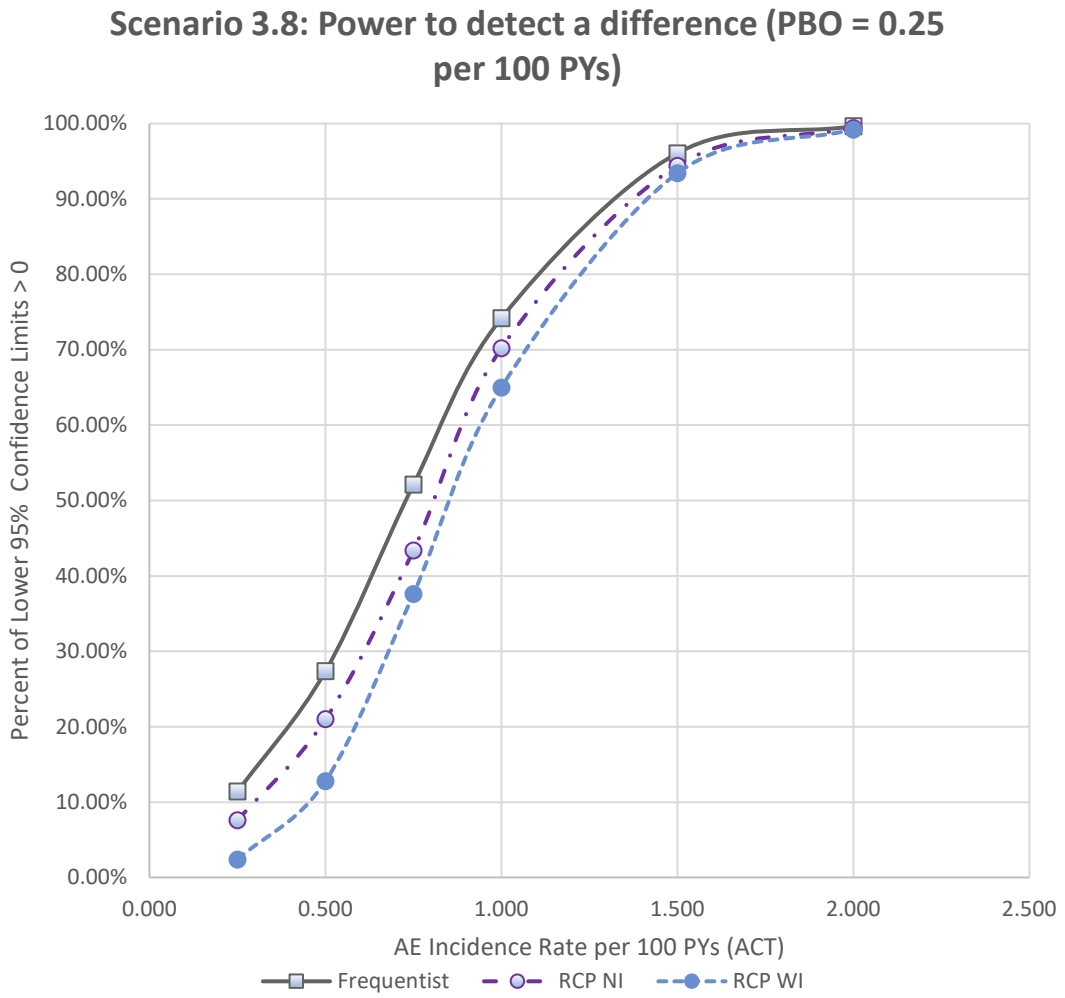
Simulation Results: Risk Difference = 0 (IR from 0.1 to 2.0 per 100 PYs)
[Scenario 3.7: Homogenous Rates Across Studies, varying RRs; N_1.=100 (1:1), N_2. and N_3.= 400 (3:1)]





Simulation Results: Risk Difference = 0 to 1.50 per 100 PYs (PBO IR = 0.25 per 100 PYs) [Scenario 3.1: Homogenous Rates Across Studies, varying RRs; N_1.=100 (1:1), N_2. and N_3.= 400 (3:1)]





Simulation Study: Summary for difference in IR

Method	Key Takeaway
Bayesian Exponential (Weakly Informative)*	**Performs satisfactorily**. Coverage a bit low,
Bayesian Right Censored Poisson (Weakly Informative)	
Bayesian Exponential (Non-informative)*	Performs poorly for small rate.
Bayesian RCP (Non-informative)	
Frequentist Asymptotic SSA	

* Based upon preliminary data

abbvie

Discussion / Conclusions

Discussion / Conclusions

- For incidence percentages, the Modified Stratified Wald, Modified Stratified t-distribution and the exact/asymptotic hybrid method are acceptable options, with the MSW and MST providing slightly better power. The asymptotic SSA Wald based method, and the Mantel-Haenszel methods produce intervals with poor coverage and subsequently unacceptable high Type I error control in specific situations (e.g. sparse events even with ample sample size). More work is needed to adequately assess the Bayesian methods.

- For incidence rates, the Bayesian Right-censored weakly informative Poisson method performs satisfactorily. The asymptotic frequentist method and the Bayesian non-informative methods not provide adequate coverage and have unacceptable Type I error rates for smaller incidence rates. However, more work is needed to adequately assess the Bayesian methods. In addition, frequentist methods (e.g., an improved Miettinen / Score type method, Hu 2023) may be viable candidates.

abbvie

References

References

[1] Crowe, B. Chuang-Stein, C. Lettis S. Brueckner, A. (2016). Reporting Adverse Drug Reactions in Product Labels. *Therapeutic Innovation & Regulatory Science*; 50(4): 455-463.

[2] Analysis and Display White Papers Project Team. (2017). Analysis and Displays Associated with Adverse Events: Focus on Adverse Events in Phase 2-4 Clinical Trials and Integrated Summary Documents. PhUSE Computational Science Standard Analyses and Code Sharing Working Group.

[3] Chuang-Stein, C. Beltangady, M. (2011). Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. *Pharmaceutical Statistics;* 10: 3-7.

[4] Sui J, Jiao J, Sun Y, Liu J, Bastero R, Koch G. (2021). Evaluation of alternative confidence intervals to address non-inferiority through the stratified difference between proportions. Pharm Stat. 2021; 20(1): 146-162.

References (cont.)

[5] Pan W. (2002). Approximate confidence intervals for one proportion and difference of two proportions. *Computational Statistics and Data Analysis;* 40: 143-157.

[6] Fagan T. (1999) Exact 95% confidence intervals for differences in binomial proportions. Comput Biol Med; 29(1): 83-7.

[7] Hong H, Wang C, Rosner GL. (2021). Meta-analysis of rare adverse events in randomized clinical trials: Bayesian and frequentist methods. Clin Trials;18(1):3-16.

[8] Hu Y., Lu K., Xie L., Zhu W. (2023). An improved score-type confidence interval for stratified risk differences involving rare events. Pharm Stat.; 22(3):492-507.

abbvie

Back UP Slides

Simpson' Paradox: Timeline and real life examples

- 1899 – Karl Pearson writes about a statistical association between two different groups is reversed when the 2 groups are combined. (Pearl, 2013)
- 1903 – Udney Yule mentions similar effects ("illusory", "fallacy might lead to seriously misleading results…", [Yule, 1903])
- 1934 – Cohen and Nagel discuss changing results in 1910 Tuberculosis death rates by race and combined populations between Richmond and NY (https://plato.stanford.edu/entries/paradox-simpson/)
- 1951 – Edward H. Simpson describes phenomenon in a technical paper[a]
- 1973 – UC Berkeley Gender Bias Graduate School Admission case[a]
- 1986 – Kidney Stone Treatment (Chiang et al, Br Med Journal)[a]
- 1996 – Appleton, French, Vanderpump. Effects of Smoking [Berman, 2012]
- 2012 – US Nationwide income change for women (2000-2012) [365 DataScience, 2017]

Solutions: Comparison of Methods

Method	Weights	Example Stratum est.	Trt Diff (2-1)
SSA	$w_i. = n_i. / (n_1. + n_2.)$	[(0.333)(0.2)+(0.667)(0.3)] – [(0.333)(0.2)+(0.667)(0.3)]	0
Cochran	$w_i. = [(n_{i1})(n_{i2})/(n_{i1} + n_{i2})]/ \sum_{i=1}^{2} [(ni_{1)(}n_{i2)/(}n_{i1+} n_{i2)}]$	[(0.308)(0.2)+(0.692)(0.3)] – [(0.308)(0.2)+(0.692)(0.3)]	0
Inverse Variance	$w_i. = [\sum_{j=1}^{2} [(p_{ij}(1-p_{ij})/n_{ij})]^{-1} / [\sum_{ij} [(p_{ij}(1-p_{ij})/n_{ij})]^{-1}$	[(0.368)(0.2)+(0.632)(0.3)] – [(0.368)(0.2)+(0.632)(0.3)]	0
Raw Pooling	$w_{ij} = n_{ij} / n._j$	[(0.400)(0.2)+(0.600)(0.3)] – [(0.250)(0.2)+(0.750)(0.3)]	-0.015

Note(1): w_{ij} = weight, n_{ij}=sample size, p_{ij} = incidence proportion; for i = trial 1 to 2, j = treatment group 1 to 2
Note(2): For example: $n_{11} = 50$, $n_{12} = 100$, $n_{21} = 150$, $n_{22} = 150$, $p_{11} = p_{12} = 0.2$, $p_{21} = p_{22} = 0.3$

Solutions: Comparison of Methods

Method	Weighting Mechanism	Pros	Cons	Other
SSA	Arithmetic means	Simple and easy to communicate / explain. Easy to extend to mult. treatment groups as well as to Inc. Rate.		Results similar to Cochran
Cochran	Harmonic means	Historical precedent . Unbiased estimator for Trt Diff.	May inflate variance	Results similar to SSA
Inverse Variance	Inverse of variance	Produce minimum variance estimators.	Breaks down for studies with 0 events. Can produce biased estimator for Trt Diff.	Weights depend upon observed event rates.
Raw Pooling	Individualized		Does not account for study (randomization not accounted for), leading to potentially biased/paradoxical outcomes	Traditional method used

Methods Investigated (Details of Each Method)

- (1) Traditional Mantel-Haenszel (MH) Weighted Method
  - The MH weighted method as applied/modified by Sato (1989) is the method used in the SAS Proc Freq procedure, when invoking the cl=MH option in the commonriskdiff statement option of the Tables statement.
  - It is provided merely as a reference point, as the MH weighting method is obviously different from the SSA weighting method
  - If all treatment groups within all studies have 0 events, a confidence interval is not produced.

- (2) Asymptotic (Wald) SSA Weighting Method
  - The classic method for computing variance for a linear combination for a discrete distribution (i.e. $\pi_1 - \pi_2$), with SSA weightings, is used to compute a standard error. The normal approximation along with the standard error estimate is used compute the 95% CI.
  - If all treatment groups within all studies have 0 events, a confidence interval is not produced.

Methods Investigated (Details of Each Method)

- (3) Asymptotic SSA Weighting with Continuity Correction
  - Applies a continuity correction to the Asymptotic SSA Weighting Method (#2). The traditional method (1/2n) which is documented in Sui et. al. (2020) and numerous other manuscripts is used for the continuity correction.

- (4) Klingenberg Method (Mantel-Haenszel weights)
  - The method proposed by Klingenberg (2013) is based upon Mantel-Haenszel weights, but improves the confidence intervals from the standard Mantel-Haenszel method (#1), by writing the variance of the Mantel-Haenszel estimator under the null of homogeneity and inverting the corresponding test.
  - It can be applied to various situations (e.g., mixture of both small and large strata size, unbalanced treatment allocations, and rare events).
  - The SAS Proc Freq procedure, with cl = Klingenberg in the commonriskdiff statement option of the Tables statement, can be used to obtain this CI.
  - If all treatment groups within all studies have 0 events, a confidence interval is not produced.

Methods Investigated (Details of Each Method)

- (5) Modified Klingenberg Method (SSA weights)
  - Replaces the Mantel-Haenszel weights with the SSA weights.
  - An additional constraint is added to prevent intractable results from occurring for extreme situations (e.g. sparse data along with higher randomization ratios)
  - This is a new proposal (not currently in the literature)

- (6) Hybrid Quasi-Exact / Asymptotic (SSA weights)
  - This is a new method proposed to combine the benefits of asymptotic based intervals along with the benefits of exact-based or quasi-exact based intervals, while minimizing the inefficiencies of each method considered independently.

Methods Investigated (Details of Each Method)

- (6) Hybrid Quasi-Exact / Asymptotic (SSA weights) (cont.)
  - It is a conditional-based approach and is defined and computed as follows:
    - (a) Compute traditional binomial asymptotic based intervals for each treatment group for each study. Subsequent steps are conditional based upon the outcomes of individual estimates.
    - (b) If all individual asymptotic based intervals (for each combination of treatment and study) produce lower confidence limits that are > 0, then use the Asymptotic (Wald) SSA Weighting Method (#2) for the integrated analysis.
    - (c) If at least one of the intervals computed in step (b) are < 0 or there are 0 events in one or more cells, then proceed as follows:
      - (c1) Re-compute the interval for each treatment group (for all affected studies) using Jeffrey's method for those cells which meet the criteria in step (c). If a cell doesn't meet the criteria in step (c), retain the asymptotic interval. Use the method from Fagan (1999) to derive the confidence interval for risk difference for each individual study, in which one or both treatment groups meet the criteria in step (c). If both treatment group don't meet the criteria in step (c), then derive the confidence interval for risk difference for the individual study using the standard asymptotic approach.

Methods Investigated (Details of Each Method)

- (6) Hybrid Quasi-Exact / Asymptotic (SSA weights) (cont.)

  - (c2) Adjust the variance estimate using a half-interval approach (and T distribution).

    For the lower interval it is: $\sigma_{i(L)} = ((RD_i - LCL_i)/ T_{(0.975,\ n_{i1} + n_{i2} -2)})^2$
    For the upper interval it is: $\sigma_{i(U)} = ((UCL_i - RD_i)/ T_{(0.975,\ n_{i1} + n_{i2} -2)})^2$

    Multiply the adjusted variance estimates by the square of the standardized SSA weights and sum across the studies to get an overall standard error estimate for each ½ interval.

  - (c3) Use the estimates obtained in step (c2) and insert in the standard asymptotic confidence interval formula to obtain the final stratified overall CI.

Methods Investigated (Details of Each Method)

- (7) Modified Stratified Wald (or T-test) Test (Sui et al, 2020)
  - Adds a fixed amount to both the numerator and denominator
  - Details are provided in Sui (2020).