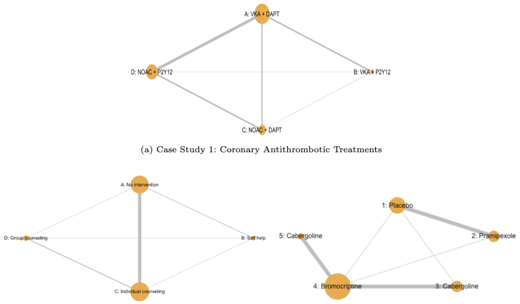
Empirical Insights into Power Behaviors of Bayesian Network Meta-Analysis: Simulation Studies and Real-World Applications

Yicheng Shen, Duke University

Sep 26, 2023

Bayesian Network Meta-analysis

- Network meta-analysis, also commonly known as the mixed treatment comparison, is an extension of the pairwise meta-analysis
- It compares multiple treatments by combining all available evidence from randomized controlled trials (RCTs) to form a network of evidence, often including similar and related studies that investigate three or more treatment arms.



(a) Case Study 1: Coronary Antithrombotic Treatments

(b) Case Study 2: Smoking Sessation Treatments

(c) Case Study 3: Parkinson Treatment

Motivations

- The power of a hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_A) is true, usually denote as $1 - \beta$.
- Powers are important, and there are lots of relevant analysis and implementation (like `G*Power` or the `pwr` R package).
- There are a few blogs, packages (`dmetar::power.analysis`) and papers discussing power analysis, but mostly in the context of MA. (Hedges and Pigott, 2001; Valentine et al., 2010; Jackson and Turner, 2017; Kruschke and Liddell, 2018).
- Power analysis in NMA can be quite challenging, restricted and complicated, but is also meaningful to its operationalizations and communications.

Method

(Generalized) Lu & Ades model

Continuous outcomes

For a network with $i = 1, \dots, I$ studies and $k = 1, \dots, K$ treatments:

$$\text{Likelihood: } \bar{y}_{ik} \sim f_Y(y_{ik} | \Delta_{ik}, \xi_{ik}) = N(\Delta_{ik}, \frac{\sigma_{ik}^2}{n_{ik}})$$

$$g(\Delta_{ik}) = \Delta_{ik} = \alpha_{iB} \qquad \text{if } k = B$$

$$g(\Delta_{ik}) = \Delta_{ik} = \alpha_{iB} + \delta_{iBk} \qquad \text{if } k \neq B$$

$$\delta_{iBk} \sim N(d_k - d_B, \tau^2)$$

(Generalized) Lu & Ades model

Binary outcomes

For a network with $i = 1, ..., I$ studies and $k = 1, ..., K$ treatments:

$$\text{Likelihood: } \bar{y}_{ik} \sim f_Y(y_{ik}|\Delta_{ik}, \xi_{ik}) = \text{Binomial}(n_{ik}, \Delta_{ik})$$

$$g(\Delta_{ik}) = \text{logit}(\Delta_{ik}) = \alpha_{iB} \qquad \text{if } k = B$$

$$g(\Delta_{ik}) = \text{logit}(\Delta_{ik}) = \alpha_{iB} + \delta_{iBk} \qquad \text{if } k \neq B$$

Hong et al. 2015

CBRE

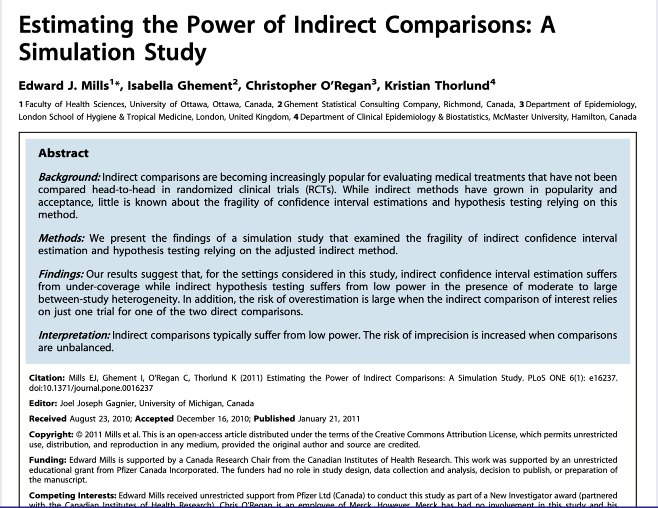
$$\begin{aligned} \text{Likelihood: } \bar{y}_{ik} &\sim f_Y(y_{ik}|\Delta_{ik}, \xi_{ik}) \\ g(\Delta_{ik}) &= \theta_{ik} = \alpha_{i1} + d_{1k} + \eta_{i1k} \\ \boldsymbol{\eta}_i &= (\eta_{i12}, ..., \eta_{i1K})^\top \sim N_{K-1}(0, \boldsymbol{\Sigma}) \end{aligned}$$

ABRE

$$\begin{aligned} \text{Likelihood: } \bar{y}_{ik} &\sim f_Y(y_{ik}|\Delta_{ik}, \xi_{ik}) \\ g(\Delta_{ik}) &= \theta_{ik} = \mu_k + \eta_{ik} \\ \boldsymbol{\eta}_i &= (\eta_{i1}, ..., \eta_{iK})^\top \sim N_K(0, \boldsymbol{\Sigma}) \\ \text{Alternatively: } \boldsymbol{\theta}_i &= (\theta_{i1}, ..., \theta_{iK})^\top \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = (\mu_1, ..., \mu_K)^\top \end{aligned}$$

Simulation study

An motivating simulation design



Estimating the Power of Indirect Comparisons: A Simulation Study

Edward J. Mills[1]*, **Isabella Ghement**[2], **Christopher O'Regan**[3], **Kristian Thorlund**[4]

1 Faculty of Health Sciences, University of Ottawa, Ottawa, Canada, **2** Ghement Statistical Consulting Company, Richmond, Canada, **3** Department of Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **4** Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Canada

Abstract

Background: Indirect comparisons are becoming increasingly popular for evaluating medical treatments that have not been compared head-to-head in randomized clinical trials (RCTs). While indirect methods have grown in popularity and acceptance, little is known about the fragility of confidence interval estimations and hypothesis testing relying on this method.

Methods: We present the findings of a simulation study that examined the fragility of indirect confidence interval estimation and hypothesis testing relying on the adjusted indirect method.

Findings: Our results suggest that, for the settings considered in this study, indirect confidence interval estimation suffers from under-coverage while indirect hypothesis testing suffers from low power in the presence of moderate to large between-study heterogeneity. In addition, the risk of overestimation is large when the indirect comparison of interest relies on just one trial for one of the two direct comparisons.

Interpretation: Indirect comparisons typically suffer from low power. The risk of imprecision is increased when comparisons are unbalanced.

Citation: Mills EJ, Ghement I, O'Regan C, Thorlund K (2011) Estimating the Power of Indirect Comparisons: A Simulation Study. PLoS ONE 6(1): e16237. doi:10.1371/journal.pone.0016237

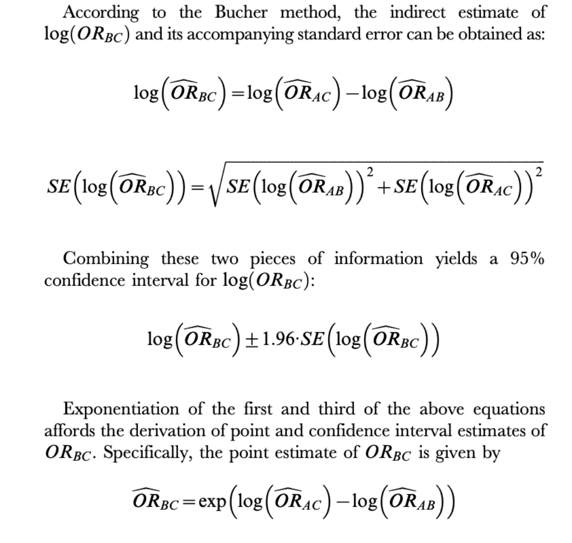
Editor: Joel Joseph Gagnier, University of Michigan, Canada

Received August 23, 2010; **Accepted** December 16, 2010; **Published** January 21, 2011

Funding: Edward Mills is supported by a Canada Research Chair from the Canadian Institutes of Health Research. This work was supported by an unrestricted educational grant from Pfizer Canada Incorporated. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Edward Mills received unrestricted support from Pfizer Ltd (Canada) to conduct this study as part of a New Investigator award (partnered with the Canadian Institutes of Health Research). Chris O'Regan is an employee of Merck. However, Merck has had no involvement in this study and his

Intuitions of computing indirect evidence from Milles et al.



According to the Bucher method, the indirect estimate of $\log(OR_{BC})$ and its accompanying standard error can be obtained as:

$$\log\left(\widehat{OR}_{BC}\right) = \log\left(\widehat{OR}_{AC}\right) - \log\left(\widehat{OR}_{AB}\right)$$

$$SE\left(\log\left(\widehat{OR}_{BC}\right)\right) = \sqrt{SE\left(\log\left(\widehat{OR}_{AB}\right)\right)^2 + SE\left(\log\left(\widehat{OR}_{AC}\right)\right)^2}$$

Combining these two pieces of information yields a 95% confidence interval for $\log(OR_{BC})$:

$$\log\left(\widehat{OR}_{BC}\right) \pm 1.96 \cdot SE\left(\log\left(\widehat{OR}_{BC}\right)\right)$$

Exponentiation of the first and third of the above equations affords the derivation of point and confidence interval estimates of OR_{BC}. Specifically, the point estimate of OR_{BC} is given by

$$\widehat{OR}_{BC} = \exp\left(\log\left(\widehat{OR}_{AC}\right) - \log\left(\widehat{OR}_{AB}\right)\right)$$

Simulation Design – Data Generation Process

$$e_{Aj} \sim Binomial\left(n_{Aj}, \pi_{Aj}\right)$$

$$e_{Bj} \sim Binomial\left(n_{Bj}, \pi_{Bj}\right)$$

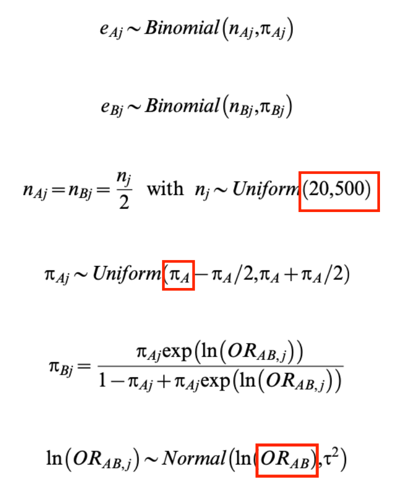
$$n_{Aj} = n_{Bj} = \frac{n_j}{2} \quad \text{with} \quad n_j \sim Uniform(20,500)$$

$$\pi_{Aj} \sim Uniform(\pi_A - \pi_A/2, \pi_A + \pi_A/2)$$

$$\pi_{Bj} = \frac{\pi_{Aj}\exp\left(\ln\left(OR_{AB,j}\right)\right)}{1 - \pi_{Aj} + \pi_{Aj}\exp\left(\ln\left(OR_{AB,j}\right)\right)}$$

$$\ln\left(OR_{AB,j}\right) \sim Normal\left(\ln(OR_{AB}), \tau^2\right)$$

Simulation Design – Data Generation Process



$$e_{Aj} \sim Binomial\left(n_{Aj}, \pi_{Aj}\right)$$

$$e_{Bj} \sim Binomial\left(n_{Bj}, \pi_{Bj}\right)$$

$$n_{Aj} = n_{Bj} = \frac{n_j}{2} \quad \text{with} \quad n_j \sim Uniform\boxed{(20,500)}$$

$$\pi_{Aj} \sim Uniform(\boxed{\pi_A} - \pi_A/2, \pi_A + \pi_A/2)$$

$$\pi_{Bj} = \frac{\pi_{Aj}\exp\left(\ln\left(OR_{AB,j}\right)\right)}{1 - \pi_{Aj} + \pi_{Aj}\exp\left(\ln\left(OR_{AB,j}\right)\right)}$$

$$\ln\left(OR_{AB,j}\right) \sim Normal\left(\ln\boxed{\left(OR_{AB}\right)}, \tau^2\right)$$

Results about powers from Milles et al.

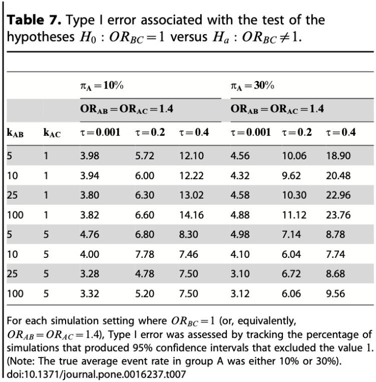


Table 7. Type I error associated with the test of the hypotheses $H_0 : OR_{BC} = 1$ versus $H_a : OR_{BC} \neq 1$.

		$\pi_A = 10\%$			$\pi_A = 30\%$		
		$OR_{AB} = OR_{AC} = 1.4$			$OR_{AB} = OR_{AC} = 1.4$		
k_{AB}	k_{AC}	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$
5	1	3.98	5.72	12.10	4.56	10.06	18.90
10	1	3.94	6.00	12.22	4.32	9.62	20.48
25	1	3.80	6.30	13.02	4.58	10.30	22.96
100	1	3.82	6.60	14.16	4.88	11.12	23.76
5	5	4.76	6.80	8.30	4.98	7.14	8.78
10	5	4.00	7.78	7.46	4.10	6.04	7.74
25	5	3.28	4.78	7.50	3.10	6.72	8.68
100	5	3.32	5.20	7.50	3.12	6.06	9.56

For each simulation setting where $OR_{BC} = 1$ (or, equivalently, $OR_{AB} = OR_{AC} = 1.4$), Type I error was assessed by tracking the percentage of simulations that produced 95% confidence intervals that excluded the value 1. (Note: The true average event rate in group A was either 10% or 30%).

doi:10.1371/journal.pone.0016237.t007

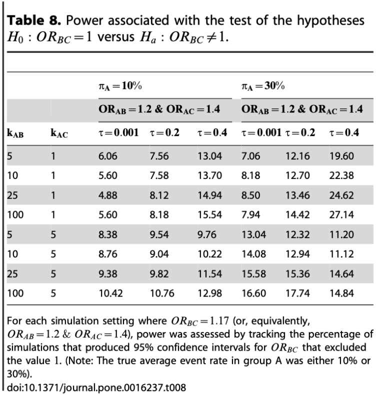


Table 8. Power associated with the test of the hypotheses $H_0 : OR_{BC} = 1$ versus $H_a : OR_{BC} \neq 1$.

		$\pi_A = 10\%$			$\pi_A = 30\%$		
		$OR_{AB} = 1.2$ & $OR_{AC} = 1.4$			$OR_{AB} = 1.2$ & $OR_{AC} = 1.4$		
k_{AB}	k_{AC}	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$
5	1	6.06	7.56	13.04	7.06	12.16	19.60
10	1	5.60	7.58	13.70	8.18	12.70	22.38
25	1	4.88	8.12	14.94	8.50	13.46	24.62
100	1	5.60	8.18	15.54	7.94	14.42	27.14
5	5	8.38	9.54	9.76	13.04	12.32	11.20
10	5	8.76	9.04	10.22	14.08	12.94	11.12
25	5	9.38	9.82	11.54	15.58	15.36	14.64
100	5	10.42	10.76	12.98	16.60	17.74	14.84

For each simulation setting where $OR_{BC} = 1.17$ (or, equivalently, $OR_{AB} = 1.2$ & $OR_{AC} = 1.4$), power was assessed by tracking the percentage of simulations that produced 95% confidence intervals for OR_{BC} that excluded the value 1. (Note: The true average event rate in group A was either 10% or 30%).

doi:10.1371/journal.pone.0016237.t008

Results about powers from Milles et al.

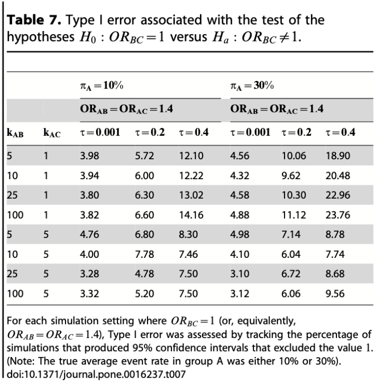


Table 7. Type I error associated with the test of the hypotheses $H_0 : OR_{BC} = 1$ versus $H_a : OR_{BC} \neq 1$.

		$\pi_A = 10\%$			$\pi_A = 30\%$		
		$OR_{AB} = OR_{AC} = 1.4$			$OR_{AB} = OR_{AC} = 1.4$		
k_{AB}	k_{AC}	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$
5	1	3.98	5.72	12.10	4.56	10.06	18.90
10	1	3.94	6.00	12.22	4.32	9.62	20.48
25	1	3.80	6.30	13.02	4.58	10.30	22.96
100	1	3.82	6.60	14.16	4.88	11.12	23.76
5	5	4.76	6.80	8.30	4.98	7.14	8.78
10	5	4.00	7.78	7.46	4.10	6.04	7.74
25	5	3.28	4.78	7.50	3.10	6.72	8.68
100	5	3.32	5.20	7.50	3.12	6.06	9.56

For each simulation setting where $OR_{BC} = 1$ (or, equivalently, $OR_{AB} = OR_{AC} = 1.4$), Type I error was assessed by tracking the percentage of simulations that produced 95% confidence intervals that excluded the value 1. (Note: The true average event rate in group A was either 10% or 30%).
doi:10.1371/journal.pone.0016237.t007

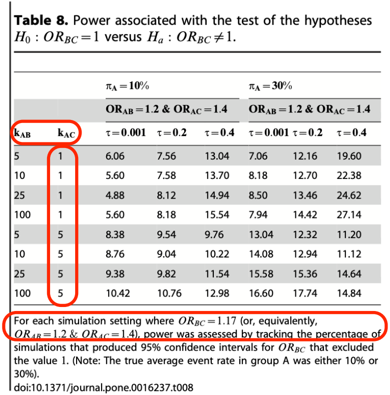


Table 8. Power associated with the test of the hypotheses $H_0 : OR_{BC} = 1$ versus $H_a : OR_{BC} \neq 1$.

		$\pi_A = 10\%$			$\pi_A = 30\%$		
		$OR_{AB} = 1.2$ & $OR_{AC} = 1.4$			$OR_{AB} = 1.2$ & $OR_{AC} = 1.4$		
k_{AB}	k_{AC}	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$	$\tau = 0.001$	$\tau = 0.2$	$\tau = 0.4$
5	1	6.06	7.56	13.04	7.06	12.16	19.60
10	1	5.60	7.58	13.70	8.18	12.70	22.38
25	1	4.88	8.12	14.94	8.50	13.46	24.62
100	1	5.60	8.18	15.54	7.94	14.42	27.14
5	5	8.38	9.54	9.76	13.04	12.32	11.20
10	5	8.76	9.04	10.22	14.08	12.94	11.12
25	5	9.38	9.82	11.54	15.58	15.36	14.64
100	5	10.42	10.76	12.98	16.60	17.74	14.84

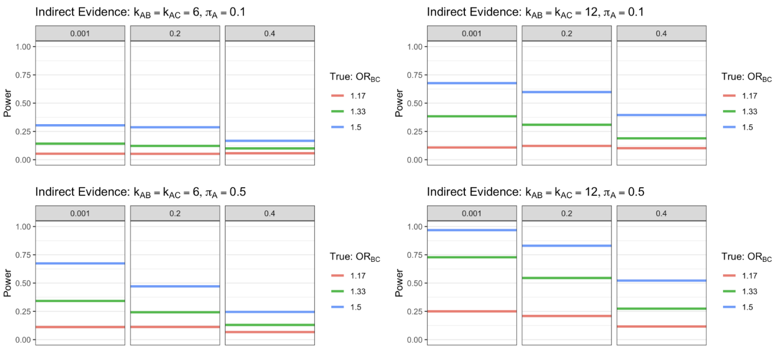
For each simulation setting where $OR_{BC} = 1.17$ (or, equivalently, $OR_{AB} = 1.2$ & $OR_{AC} = 1.4$), power was assessed by tracking the percentage of simulations that produced 95% confidence intervals for OR_{BC} that excluded the value 1. (Note: The true average event rate in group A was either 10% or 30%).
doi:10.1371/journal.pone.0016237.t008

Improvements over Miles et al. (2011)

1. We synthesize both direct and indirect evidence and fit the NMA models through Bayesian approaches instead of a frequentist one.
2. We simulate more realistic numbers of NMA studies, ranging from 1 to 6 studies of direct evidence and from 6 to 12 studies of indirect evidence.
3. We fit every simulated NMA data set with two types of aforementioned models, specifically LARE and ABRE models.
4. We examine powers of detecting more significant effect sizes that are more common in real cases, for example odds ratio of 1.17, 1.33 and 1.5.
5. We evaluate Type I and Type II error rates as well as probability of obtaining correct ranking orders of treatment effects rather than power results only.
6. We study more treatment arms and networks of different sizes and structures rather than only a three-arm network.

Results

Indirect evidence



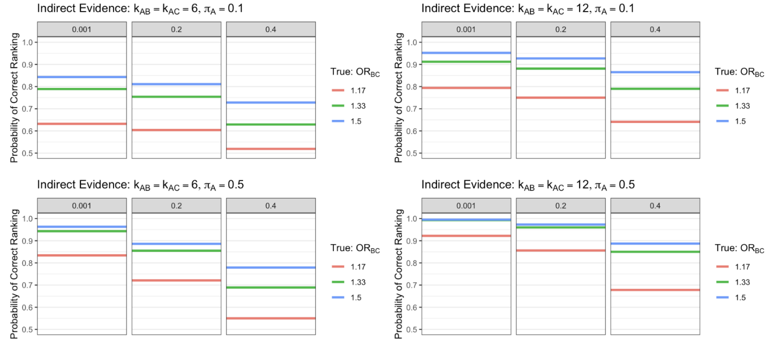
Indirect Evidence: $k_{AB} = k_{AC} = 6$, $\pi_A = 0.1$

Indirect Evidence: $k_{AB} = k_{AC} = 12$, $\pi_A = 0.1$

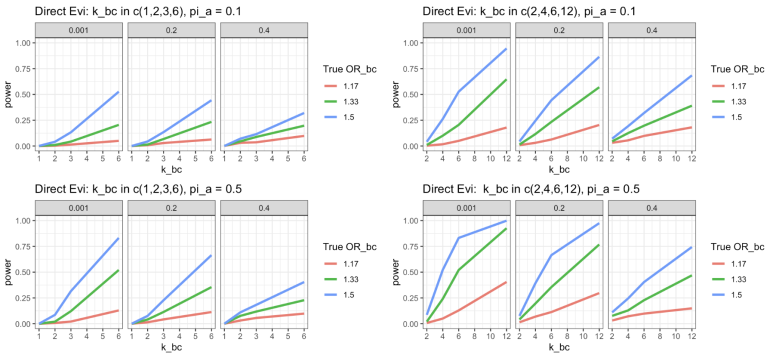
Indirect Evidence: $k_{AB} = k_{AC} = 6$, $\pi_A = 0.5$

Indirect Evidence: $k_{AB} = k_{AC} = 12$, $\pi_A = 0.5$

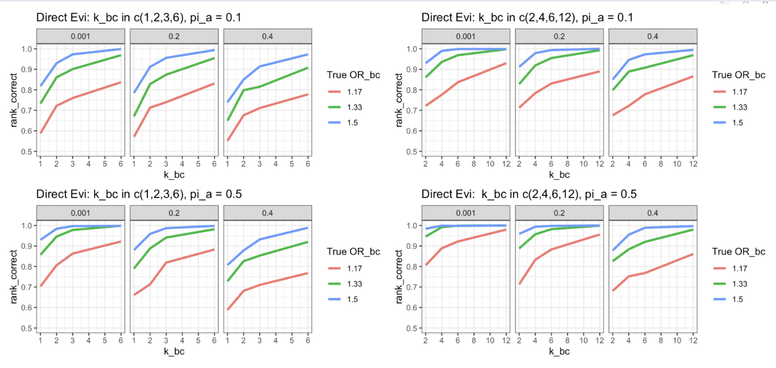
Indirect evidence



Direct evidence



Direct evidence


Direct Evi: k_bc in c(1,2,3,6), pi_a = 0.1

Direct Evi: k_bc in c(2,4,6,12), pi_a = 0.1

Direct Evi: k_bc in c(1,2,3,6), pi_a = 0.5

Direct Evi: k_bc in c(2,4,6,12), pi_a = 0.5

Indirect, direct & ovear powers in LARE

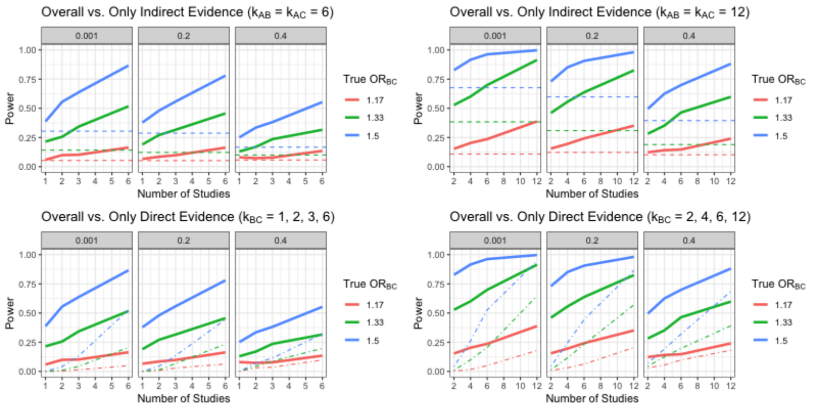


Figure 1: Power behaviors of NMA using both indirect and direct evidence. Solid lines stand for powers from overall evidence. Dashed lines stand for powers from indirect evidence. Dotdahsed lines stand for powers from direct evidence.

Indirect, direct & ovear powers in ABRE

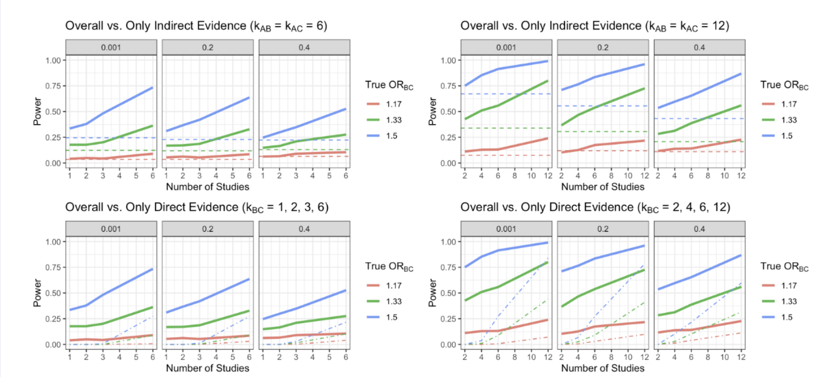
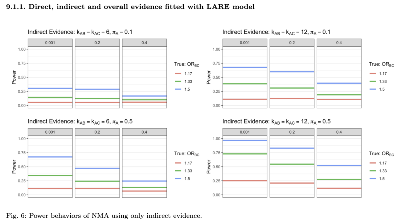


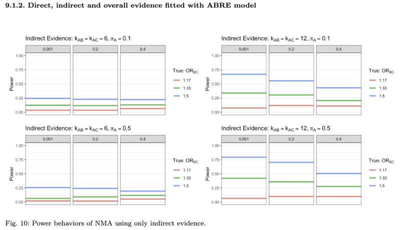
Figure 2: Power behaviors of NMA using both indirect and direct evidence. Solid lines stand for powers from overall evidence. Dashed lines stand for powers from indirect evidence. Dotdahsed lines stand for powers from direct evidence.

LARE vs. ABRE



9.1.1. Direct, indirect and overall evidence fitted with LARE model

Fig. 6: Power behaviors of NMA using only indirect evidence.



9.1.2. Direct, indirect and overall evidence fitted with ABRE model

Fig. 10: Power behaviors of NMA using only indirect evidence.

Large vs. small studies

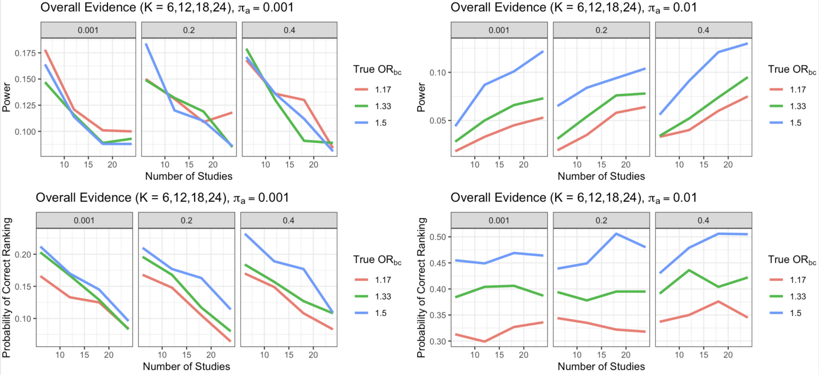
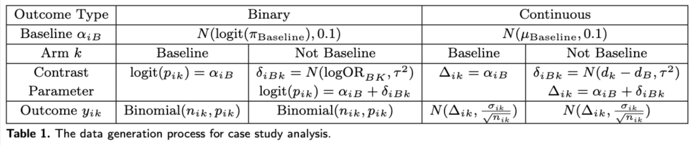


Figure 3: Power behaviors with rare and common outcomes and a total sample size of 3000.

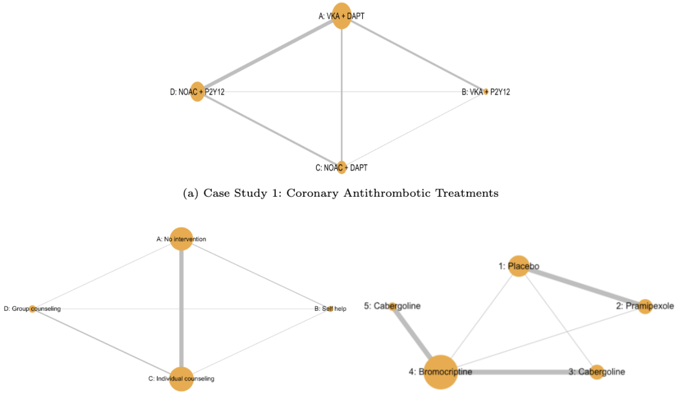
Case Studies

We considered three case studies, investigating coronary disease (Lopes et al., 2019), smoking cessation (Fiore et al., 1996) and Parkinson treatment (Franchini et al., 2012).



Outcome Type	Binary		Continuous	
Baseline α_{iB}	$N(\text{logit}(\pi_{\text{Baseline}}), 0.1)$		$N(\mu_{\text{Baseline}}, 0.1)$	
Arm k	Baseline	Not Baseline	Baseline	Not Baseline
Contrast Parameter	$\text{logit}(p_{ik}) = \alpha_{iB}$	$\delta_{iBk} = N(\text{logOR}_{BK}, \tau^2)$ $\text{logit}(p_{ik}) = \alpha_{iB} + \delta_{iBk}$	$\Delta_{ik} = \alpha_{iB}$	$\delta_{iBk} = N(d_k - d_B, \tau^2)$ $\Delta_{ik} = \alpha_{iB} + \delta_{iBk}$
Outcome y_{ik}	$\text{Binomial}(n_{ik}, p_{ik})$	$\text{Binomial}(n_{ik}, p_{ik})$	$N(\Delta_{ik}, \frac{\sigma_{ik}}{\sqrt{n_{ik}}})$	$N(\Delta_{ik}, \frac{\sigma_{ik}}{\sqrt{n_{ik}}})$

Table 1. The data generation process for case study analysis.

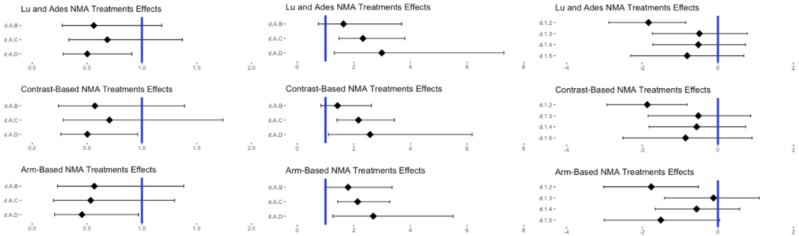


(a) Case Study 1: Coronary Antithrombotic Treatments

(b) Case Study 2: Smoking Sessation Treatments

(c) Case Study 3: Parkinson Treatment

Figure 4: Network structures of three NMA case studies: each edge represents one treatment, connecting lines indicate randomized trials directly compare pairs of treatments.

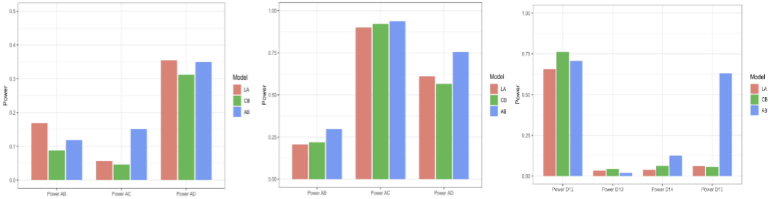


(a) Coronary Antithrombotic

(b) Smoking Sessation

(c) Parkinson Treatment

Figure 5: Point and interval estimates of NMA data under different models.



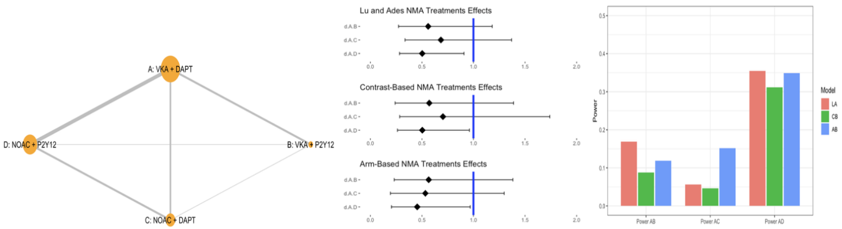
(a) Coronary Antithrombotic

(b) Smoking Sessation

(c) Parkinson Treatment

Figure 6: Power estimates from different modeling approaches.

Case Study 1



(a) Network Structure

(b) Model Estimates

(c) Simulated Powers

Fig. 3: The network structure, NMA models' point and interval estimates and power simulation results of the coronary antithrombotic case study.

Add one more study?

Adding more studies between two treatments usually greatly increases the power of detecting significant relative effects between these two treatments, if there is any. Meanwhile it can also improve our understanding of other edges in the network.

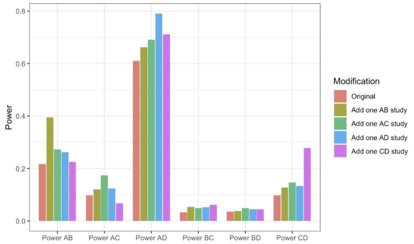
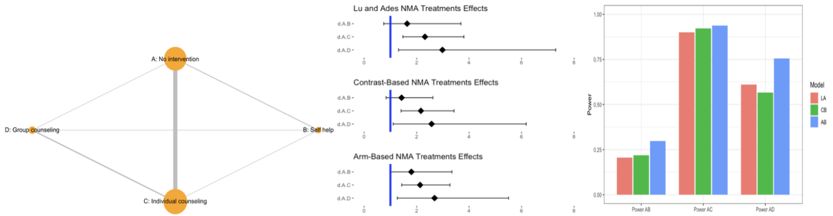


Figure 17: Power behaviors after adding one more two-arm study with average sample size ($902 \times 2 = 1804$) to the NMA using major bleeding as outcome.

Case Study 2



(a) Network Structure

(b) Model Estimates

(c) Simulated Powers

Fig. 4: The network structure, NMA models' point and interval estimates and power simulation results of the smoking cessation case study.

7.2.2 Smoking Cessation (Fiore et al., 1996)

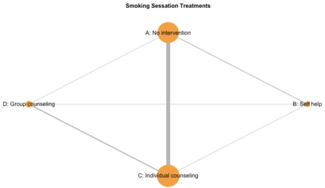


Figure 20: Network structure of four alternative smoking cessation treatments: each edge represents one treatment, connecting lines indicate randomized trials directly compare pairs of treatments.

	A	B	C	D
A		20.6	91.3	61.4
B			13.4	22.5
C				13.0
D				

(a) Powers of detecting effects via LARE model (in %)

	A	B	C	D
A		337	4860	834
B			352	383
C				894
D				

(b) Effective sample size

	A	B	C	D
A		1.53	8.71	2.91
B			1.74	1.74
C				3.39
D				

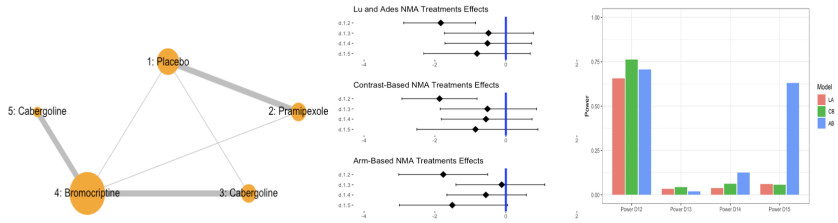
(c) Effective number of studies

	A	B	C	D
A		9.53	110.87	19.22
B			10.19	11.01
C				21.32
D				

(d) Effective precision

Table 4: Quantifying powers and effective strength of overall evidence in Case Study 2.

Case Study 3



(a) Network Structure (b) Model Estimates (c) Simulated Powers

Fig. 5: The network structure, NMA models' point and interval estimates and power simulation results of the Parkinson treatment case study.

7.2.3 Parkinson (Franchini et al., 2012)

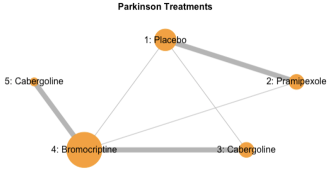

Parkinson Treatments

Figure 23: Network structure of five Parkinson treatments: each edge represents one treatment, connecting lines indicate randomized trials directly compare pairs of treatments.

	1	2	3	4	5
1		80.98	4.84	5.16	8.20

(a) Powers of detecting effects via LARE model (in %)

	1	2	3	4	5
1		588	298	375	226
2			238	322	205
3				422	242
4					565
5					

(b) Effective sample size

	1	2	3	4	5
1		2.63	1.91	2.33	1.08
2			1.40	1.91	0.98
3				2.63	1.14
4					2.00
5					

(c) Effective number of studies

	1	2	3	4	5
1		146.57	70.19	71.77	55.61
2			56.98	79.42	50.82
3				98.16	57.89
4					141.11
5					

(d) Effective precision

Table 5: Quantifying powers and effective strength of overall evidence in Case Study 3.

THANK YOU!