Method C2: Investigating potential treatment effect-modifying covariates

Method C2: Investigating potential treatment effect-modifying covariates

Including appropriate treatment by covariate interactions in NMA models.
Patient-level or trial-level covariates can be studied.
A measure of baseline risk can be explored as a trial-level covariate.

- Clinical expertise or evidence is required to determine which covariates could modify the results.
- Consider the within-trial and across-trial covariate distributions and number of contributing studies to determine if there are sufficient data.

No statistically significant regression coefficient for the interaction.

Indicates valid assumption (assuming there are no unidentified characteristics that would modify the treatment effects).

Statistically significant regression coefficient for the interaction.

Indicates invalid assumption.

Between trial variance is reduced by including the interaction.

Including appropriate treatment by covariate interactions in NMA models.
Patient-level or trial-level covariates can be studied.
A measure of baseline risk can be explored as a trial-level covariate.

- Clinical expertise or evidence is required to determine which covariates could modify the results.
- Consider the within-trial and across-trial covariate distributions and number of contributing studies to determine if there are sufficient data.

No statistically significant regression coefficient for the interaction.

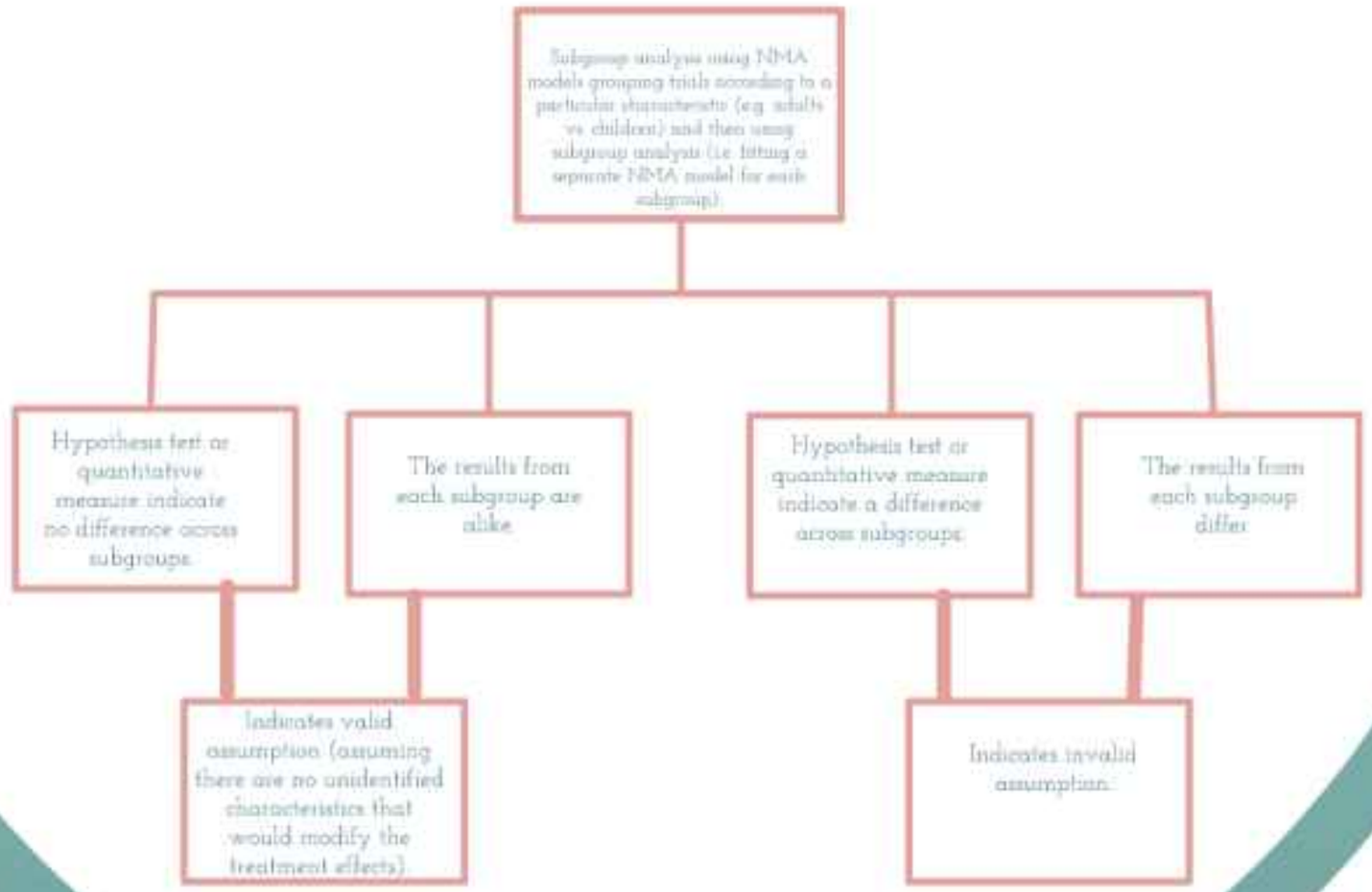
Statistically significant regression coefficient for the interaction.

Between trial variance is reduced by including the interaction.

Indicates valid assumption (assuming there are no unidentified characteristics that would modify the treatment effects).

Indicates invalid assumption.

Method C2: Investigating potential treatment effect-modifying covariates



ed to
ify the

Subgroup analysis using NMA models grouping trials according to a particular characteristic (e.g. adults vs. children) and then using subgroup analysis (i.e. fitting a separate NMA model for each subgroup).

Hypothesis test or quantitative measure indicate no difference across subgroups.

The results from each subgroup are alike.

Hypothesis test or quantitative measure indicate a difference across subgroups.

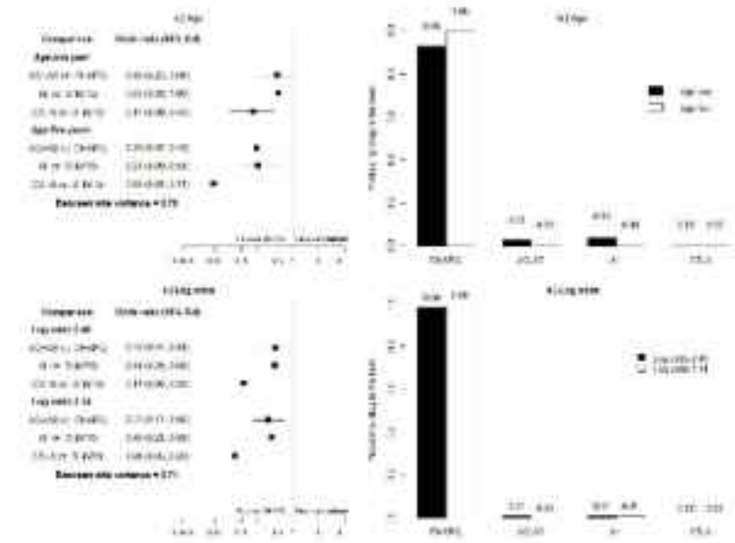
The results from each subgroup differ.

Indicates valid assumption (assuming there are no unidentified characteristics that would modify the treatment effects).

Indicates invalid assumption.

Method C2: Investigating potential treatment effect-modifying covariates

Model	Regression coefficients for treatment by covariate interactions			Between site variance
	AQ+AS vs. DHAPQ	AL vs. DHAPQ	CD+A vs. DHAPQ	
No interaction	-	-	-	0.71 (0.26, 1.86)
Treatment by age interaction	-0.23 (-0.47, 0.01)	-0.23 (-0.46, 0.00)	-0.39 (-0.74, -0.09)	0.79 (0.28, 2.09)
Treatment by log odds interaction	-0.50 (-1.22, 0.27)	-0.15 (-0.82, 0.55)	-0.57 (-1.58, 0.28)	0.71 (0.25, 1.79)



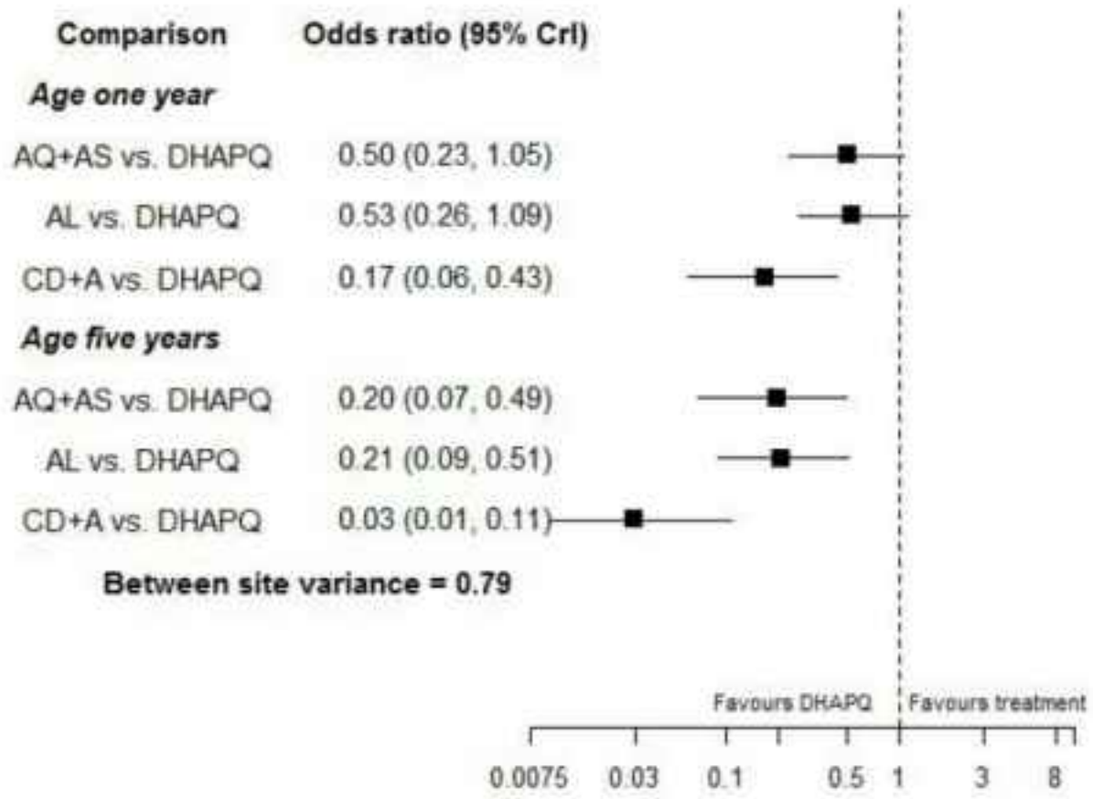
Variance higher with age interaction.

Model	Regression coefficients for treatment by covariate interactions			Between site variance
	AQ+AS vs. DHAPQ	AL vs. DHAPQ	CD+A vs. DHAPQ	
No interaction	-	-	-	0.71 (0.26, 1.86)
Treatment by age interaction	-0.23 (-0.47, 0.01)	-0.23 (-0.46, 0.00)	-0.38 (-0.74, -0.09)	0.79 (0.28, 2.09)
Treatment by log odds interaction	-0.50 (-1.22, 0.27)	-0.15 (-0.82, 0.55)	-0.57 (-1.58, 0.28)	0.71 (0.25, 1.79)

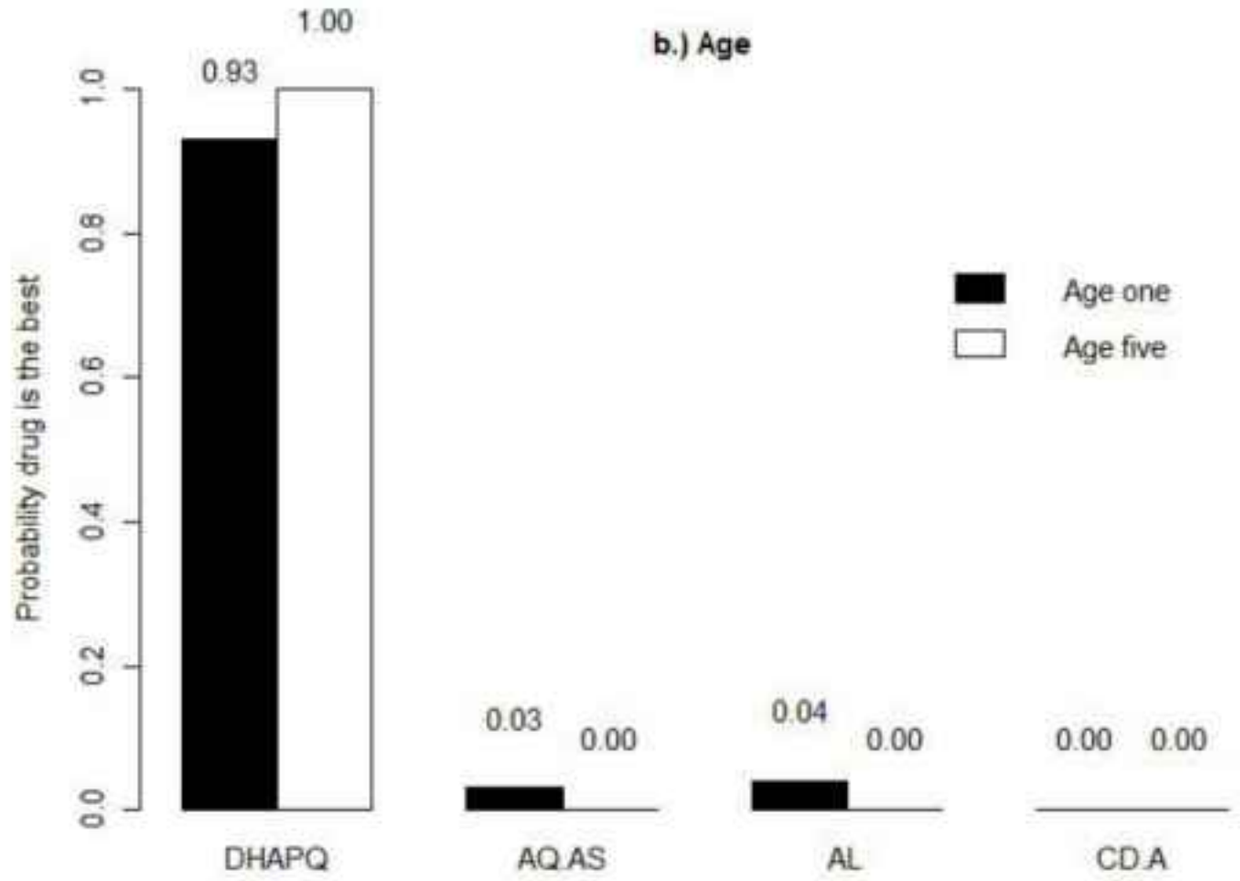
Significant interaction
indicating that an interaction exists.

Assumption is dubious because interactions exist.

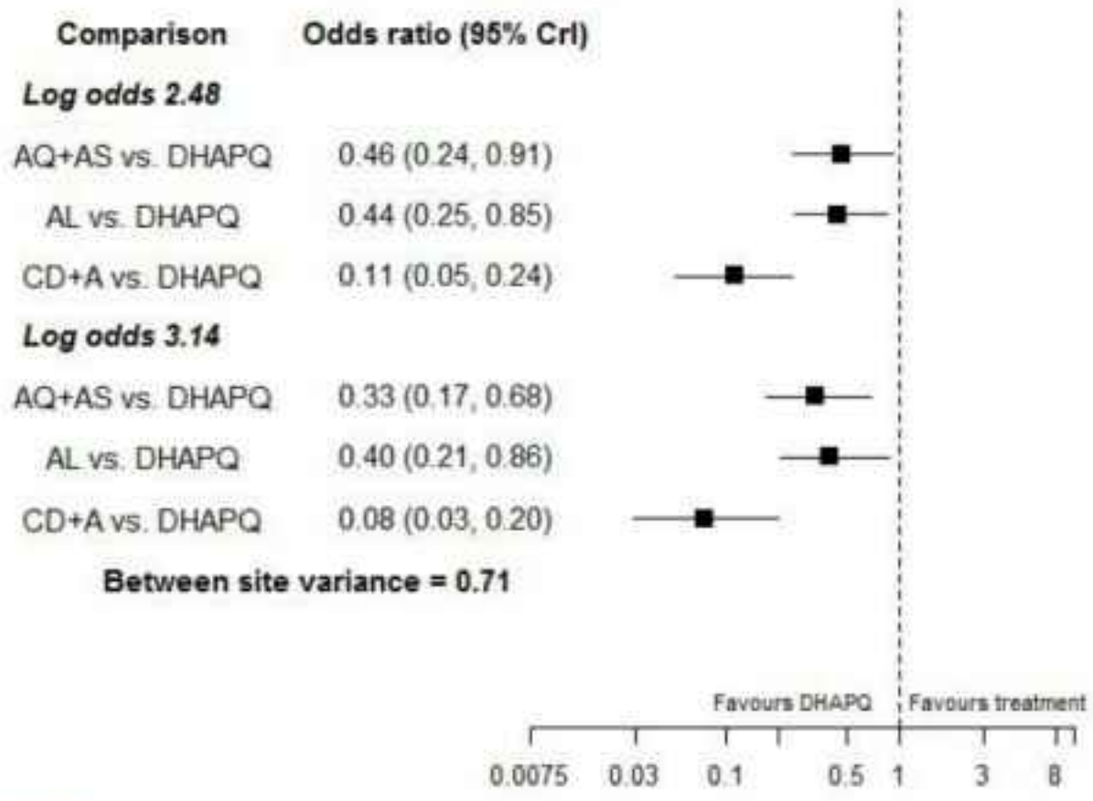
a.) Age



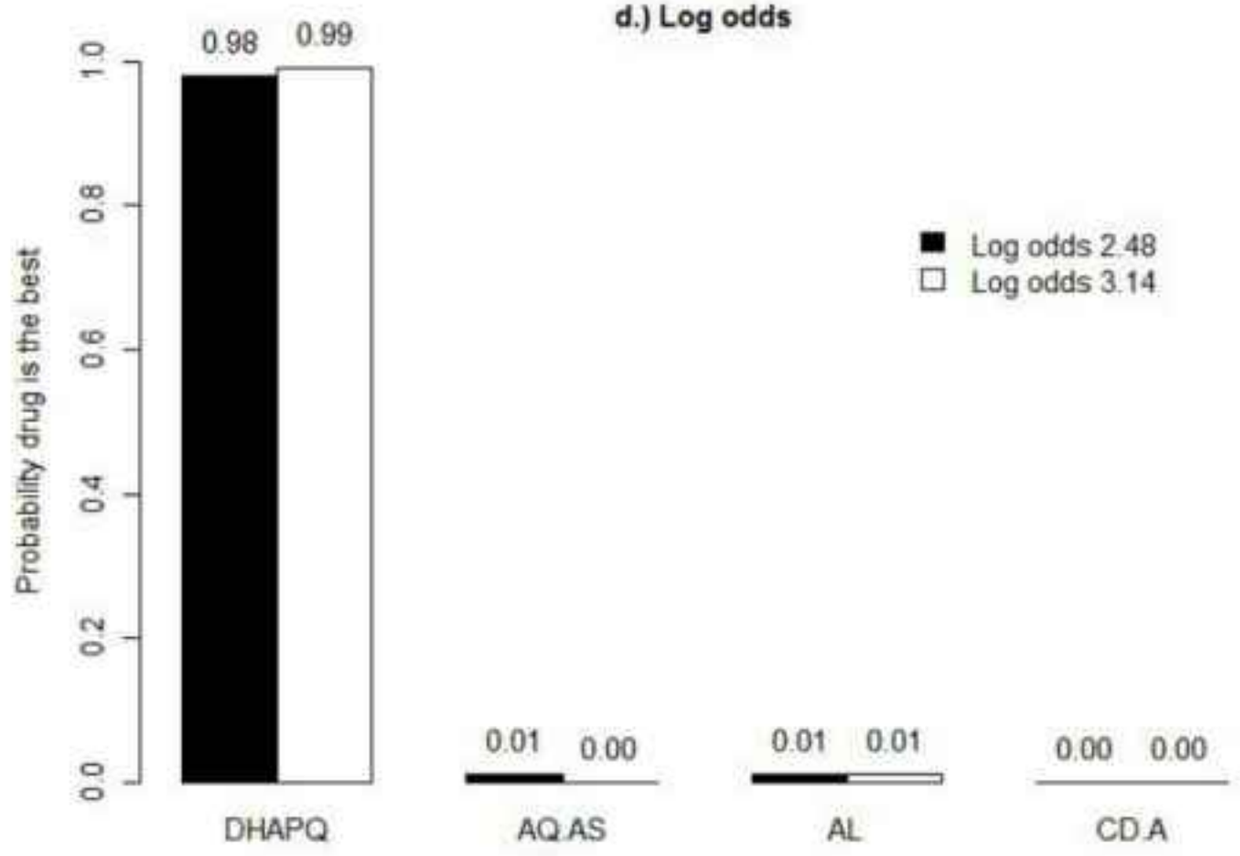
b.) Age



c.) Log odds

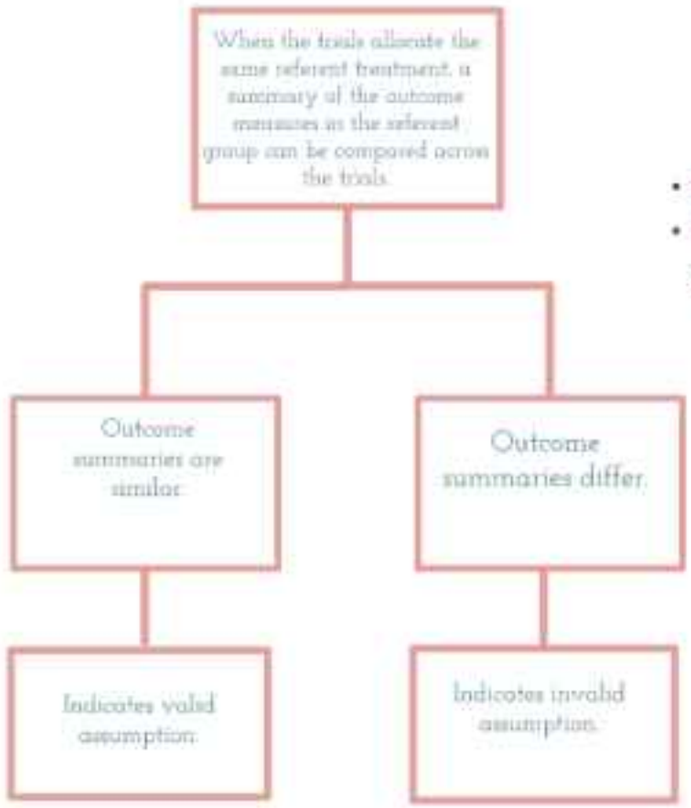


d.) Log odds



Method C3: Comparing outcome measurements in the referent group

Method C3: Comparing outcome measurements in the referent group



- Subjective
- All trials must allocate the same referent treatment.

When the trials allocate the same referent treatment, a summary of the outcome measures in the referent group can be compared across the trials.

- Subjective
- All trials must allocate t
  same referent treatmen

Outcome summaries are similar.

Outcome summaries differ.

Indicates valid assumption.

Indicates invalid assumption.

Method C3: Comparing outcome measurements in the referent group

African site	Dihydroartemisinin Piperaquine (DHAPQ)
Manhica (after CD+A)	94/100
Mbarara (after CD+A)	63/65
Nanoro	187/219
Gabon	62/63
Afokang	67/72
Pamol	60/65
Ndola	67/67
Manhica (before CD+A)	78/82
Mbarara (before CD+A)	72/80
Rukara (after CD+A)	46/47
Jinja (after CD+A)	160/167
Tororo (after CD+A)	54/75
Mashesha (after CD+A)	49/52
Rukara (before CD+A)	22/23
Jinja (before CD+A)	37/39
Tororo (before CD+A)	109/141
Mashesha (before CD+A)	23/24

Treatment success rates for patients given DHAPQ varied from 72% to 100% across sites

The variability may reflect differences in reinfection rates and varying transmission rates.

These differences could modify the treatment effects and violate the assumption.

African site	Dihydroartemisinin-Piperaquine (DHAPQ)
Manhica (after CD+A)	94/100
Mbarara (after CD+A)	63/65
Nanoro	187/219
Gabon	62/63
Afokang	67/72
Pamol	60/65
Ndola	67/67
Manhica (before CD+A)	78/82
Mbarara (before CD+A)	72/80
Rukara (after CD+A)	46/47
Jinja (after CD+A)	160/167
Tororo (after CD+A)	54/75
Mashesha (after CD+A)	49/52
Rukara (before CD+A)	22/23
Jinja (before CD+A)	37/39
Tororo (before CD+A)	109/141
Mashesha (before CD+A)	23/24

Treatment success rates for patients given DHAPQ varied from 72% to 100% across sites

The variability may reflect differences in reinfection rates and varying transmission rates.

These differences could modify the treatment effects and violate the assumption.

Method C4: Node-splitting (side splitting)

Method C4: Node-Splitting (Side Splitting)

- Visual interpretation is subjective.
- Can only be applied when both direct and indirect evidence exists for a particular comparison.

Applying a series of node-splitting NMA models (one model for each treatment pairing for which there is direct and indirect evidence).
A node-splitting model for a particular treatment comparison k vs b, splits the treatment effect of k vs b within the model, to enable an estimate of the treatment effect of k vs b using direct evidence, and an estimate of the treatment effect of k vs b from all indirect evidence, to be obtained from that particular model.

Treatment effects estimated from direct and indirect evidence are similar (using Bayesian P value and visual interpretation)

The NMA model without node-splitting is a better fitting model

Indicates valid assumption

Treatment effects estimated from direct and indirect evidence differ (using Bayesian P value and visual interpretation)

One or more node-splitting model is a better fitting model

Indicates invalid assumption

erpretation is
ective.
plied when both
direct evidence
a particular
parison.

Applying a series of node-splitting NMA models (one model for each treatment pairing for which there is direct and indirect evidence).
A node-splitting model for a particular treatment comparison k vs. b, splits the treatment effect of k vs. b within the model, to enable an estimate of the treatment effect of k vs. b using direct evidence, and an estimate of the treatment effect of k vs. b from all indirect evidence, to be obtained from that particular model.

Treatment effects estimated from direct and indirect evidence are similar (using Bayesian P value and visual interpretation).

The NMA model without node-splitting is a better fitting model.

Indicates valid assumption.

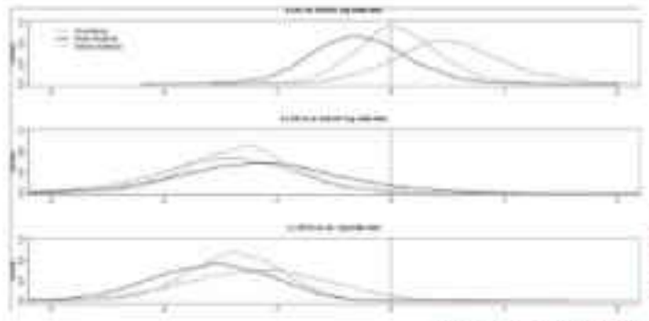
Treatment effects estimated from direct and indirect evidence differ (using Bayesian P value and visual interpretation).

One or more node-splitting model is a better fitting model.

Indicates invalid assumption.

Method C4: Node-Splitting (Side Splitting)

Model	Number of sites (number of patients)		
	All evidence	Direct evidence	Indirect evidence
Model without node-splitting	17 (3874)	-	-
AQ+AS vs. DHAPQ	-	9 (1742)	-
AL vs. DHAPQ	-	13 (2225)	-
CD+A vs. DHAPQ	-	6 (782)	-
Node-split: AL vs. AQ+AS	-	5 (1201)	17 (2877)
Node-split: CD+A vs. AQ+AS	-	2 (329)	13 (2196)
Node-split: CD+A vs. AL	-	4 (449)	15 (2551)

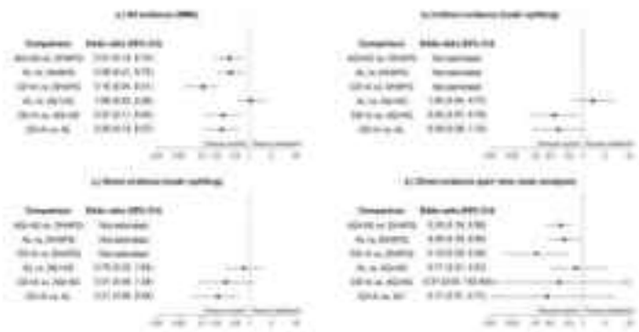


AL vs. AQ+AS: P=0.2
CD+A vs. AQ+AS: p=0.76
CD+A vs. AL: P=0.64

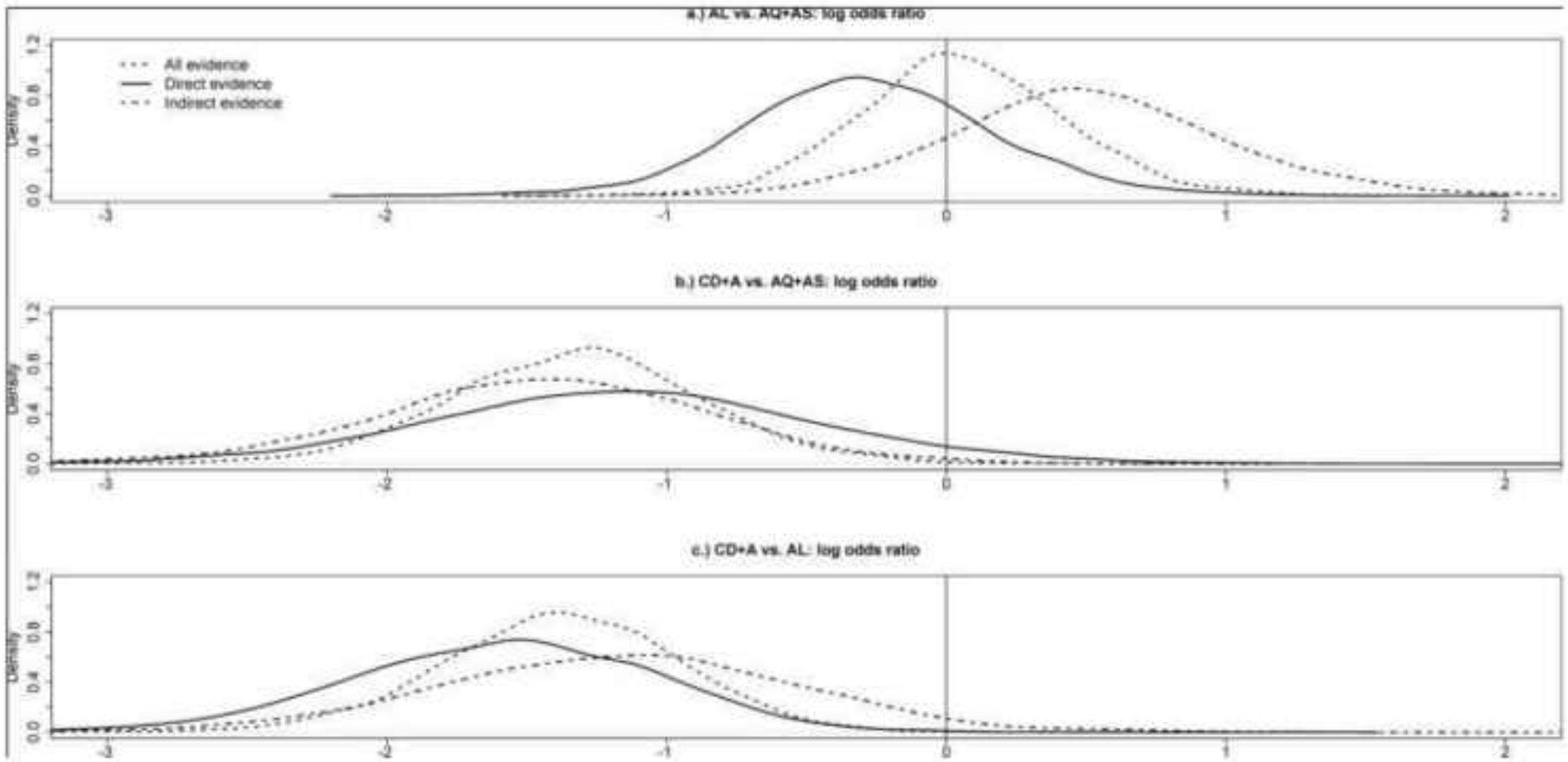
No statistically significant differences between direct and indirect evidence are detected (P>0.1).

Model	Residual deviance	Number of model parameters	Number of effective parameters	DIC	Between site variance (median (95% CrI))
Model without node-splitting	3183.0	21	35.7	3218.7	0.71 (0.26, 1.86)
Node-split: AL vs. AQ+AS	3182.0	22	35.4	3217.4	0.61 (0.21, 1.67) (14% decrease)
Node-split: CD+A vs. AQ+AS	3183.0	22	36.2	3219.2	0.78 (0.28, 2.00) (7% increase)
Node-split: CD+A vs. AL	3183.0	22	35.8	3218.8	0.69 (0.26, 1.80) (3% decrease)

The DIC when each node is split are similar to the DIC from the NMA model without node-splitting, indicating that there is no obviously superior model.



Model	Number of sites (number of patients)		
	All evidence	Direct evidence	Indirect evidence
Model without node-splitting	17 (3874)	-	-
AQ+AS vs. DHAPQ	-	9 (1742)	-
AL vs. DHAPQ	-	13 (2225)	-
CD+A vs. DHAPQ	-	6 (782)	-
Node-split: AL vs. AQ+AS	-	5 (1201)	17 (2877)
Node-split: CD+A vs. AQ+AS	-	2 (329)	13 (2198)
Node-split: CD+A vs. AL	-	4 (449)	15 (2551)



For AL vs. AQ+AS, the direct evidence does not exactly coincide with the indirect evidence.

For CDA vs AQ+AS and CD+A vs AL, distributions generally overlap indicating there is reasonably good agreement between direct and indirect evidence.

AL vs. AQ+AS: P=0.2

CD+A vs. AQ+AS: p=0.76

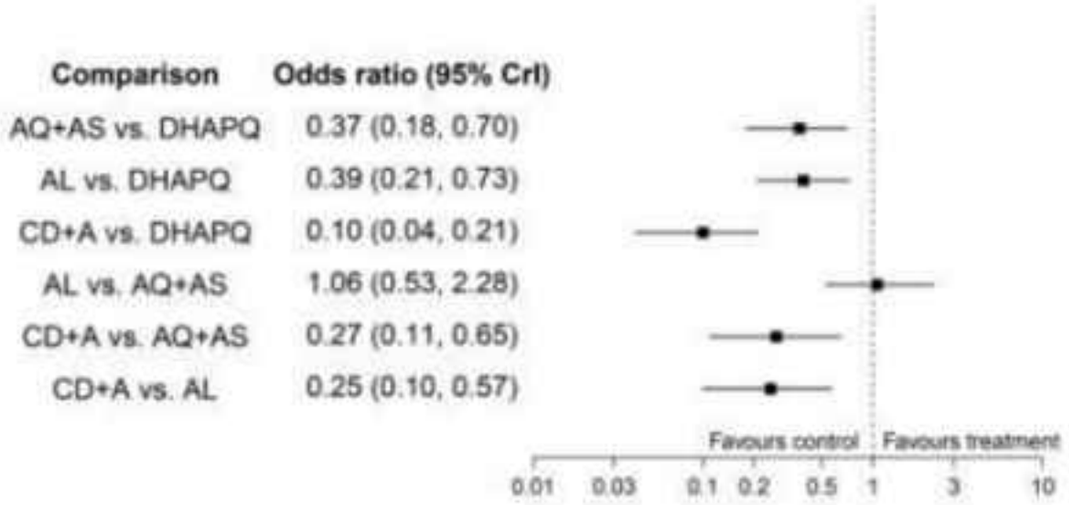
CD+A vs. AL: P=0.64

No statistically significant differences between direct and indirect evidence are detected (P>0.1).

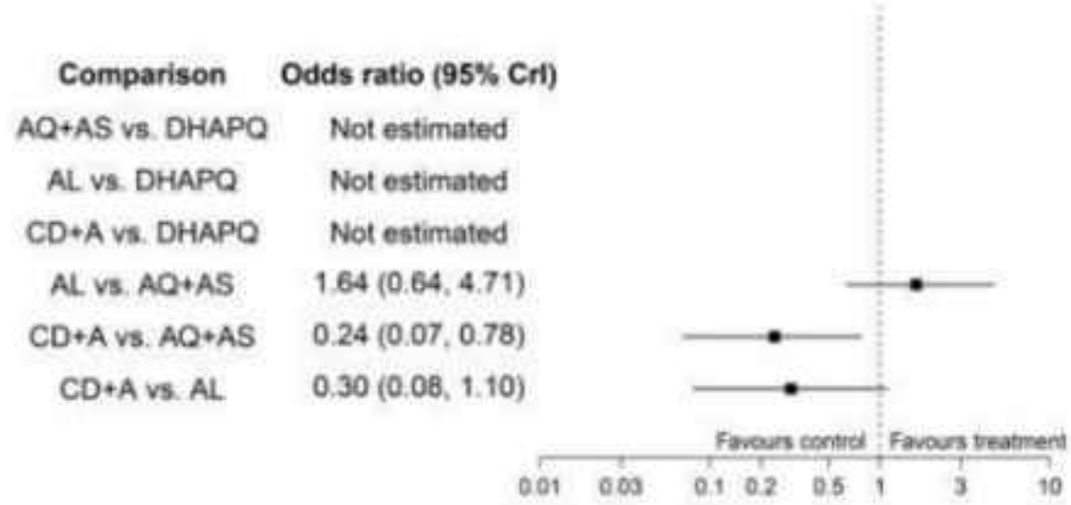
Model	Residual deviance	Number of model parameters	Number of effective parameters	DIC	Between site variance (median (95% CrI))
Model without node-splitting	3183.0	21	35.7	3218.7	0.71 (0.26, 1.86)
Node-split: AL vs. AQ+AS	3182.0	22	35.4	3217.4	0.61 (0.21, 1.67) (14% decrease)
Node-split: CD+A vs. AQ+AS	3183.0	22	36.2	3219.2	0.76 (0.28, 2.00) (7% increase)
Node-split: CD+A vs. AL	3183.0	22	35.8	3218.8	0.69 (0.26, 1.80) (3% decrease)

The DIC when each node is split are similar to the DIC from the NMA model without node-splitting, indicating that there is no obviously superior model.

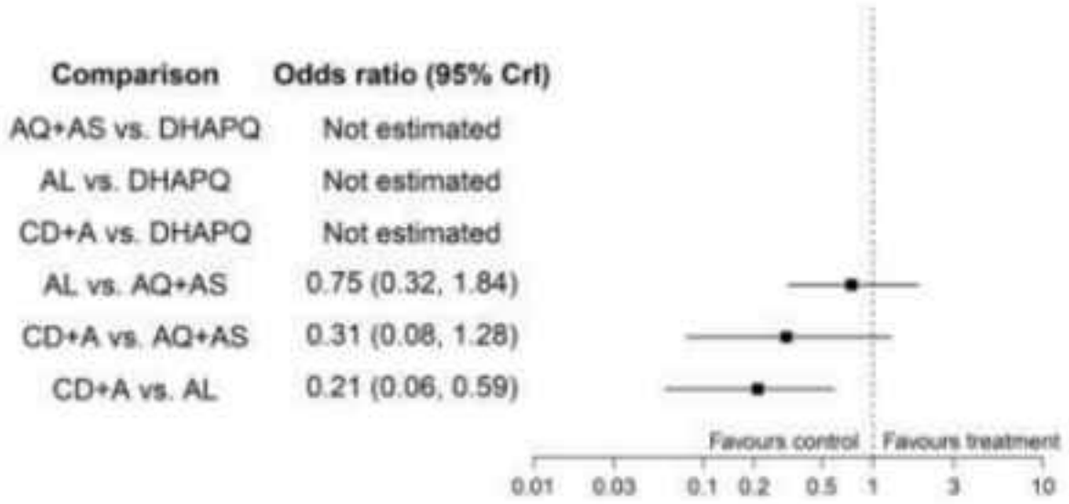
a.) All evidence (NMA)



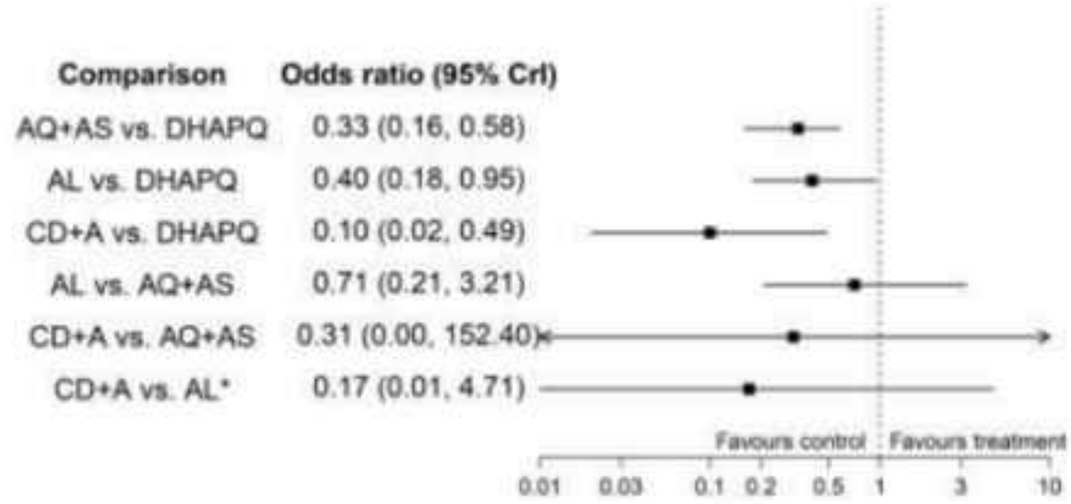
b.) Indirect evidence (node-splitting)



c.) Direct evidence (node-splitting)



d.) Direct evidence (pair-wise meta-analyses)



For CD+A versus AQ+AS and CD+A versus AL, results from each evidence type are agreeable.
For AL versus AQ+AS, the direct evidence and indirect evidence is conflicting; the direct evidence favours AQ+AS, but the indirect evidence favours AL, although neither result is significant.
Assumption is debateable.

Method C5: Inconsistency models

Inconsistency models (method C5)

Including inconsistency parameters in the NMA model (the Lu and Ades model, the design by treatment model, the Lumley model, the two-way model, and the pair-wise model).

The inconsistency model is a better fitting model

No statistically significant or large inconsistency parameters

Wald test does not detect inconsistency

The inconsistency model is a better fitting model

Statistically significant or large inconsistency parameters

Wald test detects inconsistency

Indicates valid assumption

Indicates invalid assumption

- Multi-arm trials may preclude the usefulness of some proposed models.
- Treating inconsistency parameters as random-effects when the number of inconsistency parameters is low or when the number of studies contributing to the estimation of an inconsistency parameters is small, results in estimates of variance and inconsistency parameters that are influenced by the prior distribution.

Inconsistency models (method C5)

Including inconsistency parameters in the NMA model (the Lu and Ades model, the design by treatment model, the Lumley model, the two-way model, and the pair-wise model).

The consistency model is a better fitting model.

No statistically significant or large inconsistency parameters

Wald test does not detect inconsistency

The inconsistency model is a better fitting model.

Statistically significant or large inconsistency parameters

Wald test detects inconsistency

Indicates valid assumption.

Indicates invalid assumption.

- Multi-arm trials may preclude the usefulness of some proposed models.
- Treating inconsistency parameters as random-effects when the number of inconsistency parameters is low or when the number of studies contributing to the estimation of an inconsistency parameters is small, results in estimates of variance and inconsistency parameters

cates valid
sumption.

Indicates invalid
assumption.

- Multi-arm trials may preclude the usefulness of some proposed models.
- Treating inconsistency parameters as random-effects when the number of inconsistency parameters is low or when the number of studies contributing to the estimation of an inconsistency parameters is small, results in estimates of variance and inconsistency parameters that are influenced by the prior distribution.

Inconsistency models (method C5)

Design group	Treatments	A	B	C	D
1	DHAPQ (A), AQ+AS (B).	Referent	BAB	-	-
2	DHAPQ (A), AQ+AS (B), AL (C).	Referent	BAB + ω2#B	BAC	-
3	DHAPQ (A), AQ+AS (B), CD+A (D).	Referent	BAB + ω3#B	-	BAD
4	DHAPQ (A), AL (C).	Referent	-	BAC + ω4#C	-
5	DHAPQ (A), AL (C), CD+A (D).	Referent	-	BAC + ω5AC	BAD+ ω5AD

Result	Consistency model	Inconsistency model
Residual deviance	3183.0	3183.0
Number of model parameters1	21	26
Number of effective parameters	35.4	37.7
DIC	3218.4	3220.7

Wald test for consistency:
test statistic= 1.97, df=5, P=0.85
Statistically significant inconsistency is not detected.

The DIC for the inconsistency model is similar to that of the consistency model suggesting that no model is clearly superior in terms of model fit and complexity

No inconsistency parameters are statistically significant.

Some design inconsistency within the network which is most likely caused by varying resistance patterns and transmission rates across sites.

Result		Consistency model	Inconsistency model
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ	-0.89 (-1.55, -0.27)	-
	AL vs. DHAPQ	-0.84 (-1.40, -0.21)	-
	CD+A vs. DHAPQ	-2.17 (-2.97, -1.43)	-
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS} design	-	-1.41 (-3.00, 0.17)
	AL vs. DHAPQ in {DHAPQ, AQ+AS, AL} design	-	-1.14 (-2.15, -0.10)
	CD+A vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design	-	-2.08 (-3.52, -0.57)
Inconsistency parameter (median (95% CrI); mean (standard error))	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, AL} design compared with {DHAPQ, AQ+AS} design	-	0.57 (-1.35, 2.45); 0.57 (0.95)
	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design compared with {DHAPQ, AQ+AS} design	-	0.46 (-1.70, 2.64); 0.46 (1.09)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL} design compared with {DHAPQ, AQ+AS, AL} design	-	0.59 (-0.96, 2.17); 0.59 (0.78)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, AL} design	-	0.82 (-0.76, 2.49); 0.82 (0.82)
	Design inconsistency for CD+A vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, CD+A} design	-	0.09 (-1.87, 1.94); 0.08 (0.96)
Between site variance (median (95% CrI))		0.66 (0.25, 1.71)	0.85 (0.29, 2.34)

Inconsistency parameters are smaller than their standard errors except for one where inconsistency parameter equal to its standard error which may reflect inconsistency in the effect between the two designs.

The results (log odds ratio (95% CrI)) for AL vs. DHAPQ estimated from the {DHAPQ, AL, CD+A} design (i.e. -0.32 (-1.54, 1.01)) differ from the results estimated from the {DHAPQ, AQ+AS, AL} design (i.e. -1.14 (-2.15, -0.10)).

Design group	Treatments	A	B	C	D
1	DHAPQ (*A*), AQ+AS (*B*).	Referent	*θAB*	-	-
2	DHAPQ (*A*), AQ+AS (*B*), AL (*C*).	Referent	*θAB* + ω2*AB*	*θAC*	-
3	DHAPQ (*A*), AQ+AS (*B*), CD+A (*D*).	Referent	*θAB* + ω3*AB*	-	*θAD*
4	DHAPQ (*A*), AL (*C*).	Referent	-	*θAC* + ω4*AC*	-
5	DHAPQ (*A*), AL (*C*), CD+A (*D*).	Referent	-	*θAC* + ω5*AC*	*θAD*+ ω5*AD*

Wald test for consistency:

method C5)

Result	Consistency model	Inconsistency model
Residual deviance	3183.0	3183.0
Number of model parameters1	21	26
Number of effective parameters	35.4	37.7
DIC	3218.4	3220.7

The DIC for the inconsistency model is similar to that of the consistency model suggesting that no model is clearly superior in terms of model fit and complexity

t	Consistency model	Inconsistency model
AS vs. DHAPQ	-0.89 (-1.55, -0.27)	-
PAPQ	-0.84 (-1.40, -0.21)	-

Inconsistency parameters are smaller than their standard errors except for one where

	DHAPQ (*A*), AQ+AS (*B*), CD+A (*D*).	Referent	ω3*AB*	-	ω*AD*
4	DHAPQ (*A*), AL (*C*).	Referent	-	*θAC* + ω4*AC*	-
5	DHAPQ (*A*), AL (*C*), CD+A (*D*).	Referent	-	*θAC* + ω5*AC*	*θAD*+ ω5*AD*

Wald test for consistency:

test statistic= 1.97, df=5, P=0.85

Statistically significant inconsistency is not detected.

Log odds ratio (median (95% CrI))

Log odds ratio

and complexity

	Result	Consistency model	Inconsistency model
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ	-0.89 (-1.55, -0.27)	-
	AL vs. DHAPQ	-0.84 (-1.40, -0.21)	-
	CD+A vs. DHAPQ	-2.17 (-2.97, -1.43)	-
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS} design	-	-1.41 (-3.00, 0.17)
	AL vs. DHAPQ in {DHAPQ, AQ+AS, AL} design	-	-1.14 (-2.15, -0.10)
	CD+A vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design	-	-2.08 (-3.52, -0.57)
Inconsistency parameter (median (95% CrI); mean (standard error))	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, AL} design compared with {DHAPQ, AQ+AS} design	-	0.57 (-1.35, 2.45); 0.57 (0.95)
	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design compared with {DHAPQ, AQ+AS} design	-	0.46 (-1.70, 2.64); 0.46 (1.09)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL} design compared with {DHAPQ, AQ+AS, AL} design	-	0.59 (-0.96, 2.17); 0.59 (0.78)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, AL} design	-	0.82 (-0.76, 2.49); 0.82 (0.82)
	Design inconsistency for CD+A vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, CD+A} design	-	0.09 (-1.87, 1.94); 0.08 (0.96)
Between site variance (median (95% CrI))		0.66 (0.25, 1.71)	0.85 (0.29, 2.34)

	AQ+AS (*B*), CD+A (*D*).		ω3*AB*		
4	DHAPQ (*A*), AL (*C*).	Referent	-	θ*AC* + ω4*AC*	-
5	DHAPQ (*A*), AL (*C*), CD+A (*D*).	Referent	-	θ*AC* + ω5*AC*	θ*AD*+ ω5*AD*

Number of model parameters	21	20
Number of effective parameters	35.4	37.7
DIC	3218.4	3220.7

Wald test for consistency:
test statistic= 1.97, df=5, P=0.85
Statistically significant inconsistency is not detected.

The DIC for the inconsistency model is similar to that of the consistency model suggesting that no model is clearly superior in terms of model fit and complexity

No inconsistency parameters are statistically significant.

Some design inconsistency within the network which is most likely caused by varying resistance patterns and transmission rates across sites.

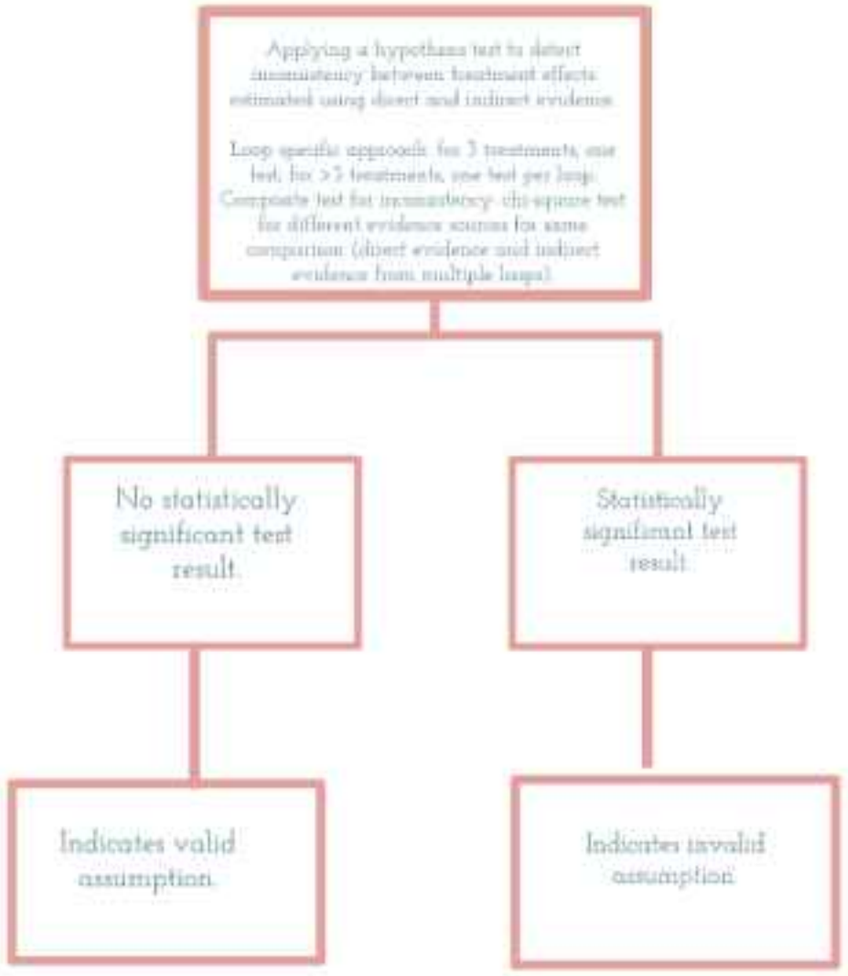
	Result	Consistency model	Inconsistency model
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ	-0.89 (-1.55, -0.27)	-
	AL vs. DHAPQ	-0.84 (-1.40, -0.21)	-
	CD+A vs. DHAPQ	-2.17 (-2.97, -1.43)	-
Log odds ratio (median (95% CrI))	AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS} design	-	-1.41 (-3.00, 0.17)
	AL vs. DHAPQ in {DHAPQ, AQ+AS, AL} design	-	-1.14 (-2.15, -0.10)
	CD+A vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design	-	-2.08 (-3.52, -0.57)
Inconsistency parameter (median (95% CrI); mean (standard error))	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, AL} design compared with {DHAPQ, AQ+AS} design	-	0.57 (-1.35, 2.45); 0.57 (0.95)
	Design inconsistency for AQ+AS vs. DHAPQ in {DHAPQ, AQ+AS, CD+A} design compared with {DHAPQ, AQ+AS} design	-	0.46 (-1.70, 2.64); 0.46 (1.09)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL} design compared with {DHAPQ, AQ+AS, AL} design	-	0.59 (-0.96, 2.17); 0.59 (0.78)
	Design inconsistency for AL vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, AL} design	-	0.82 (-0.76, 2.49); 0.82 (0.82)
	Design inconsistency for CD+A vs. DHAPQ in {DHAPQ, AL, CD+A} design compared with {DHAPQ, AQ+AS, CD+A} design	-	0.09 (-1.87, 1.94); 0.08 (0.96)
Between site variance (median (95% CrI))		0.66 (0.25, 1.71)	0.85 (0.29, 2.34)

Inconsistency parameters are smaller than their standard errors except for one where inconsistency parameter equal to its standard error which may reflect inconsistency in the effect between the two designs.

The results (log odds ratio (95% CrI)) for AL vs. DHAPQ estimated from the {DHAPQ, AL, CD+A} design (i.e. -0.32 (-1.54, 1.01)) differ from the results estimated from the {DHAPQ, AQ+AS, AL} design (i.e. -1.14 (-2.15, -0.10)).

Method C6: Applying a hypothesis test

Applying a hypothesis test (method C6)



- Multi-arm trials preclude the usefulness of the method
- Results are influenced by the number of evidence sources or number of studies.

Applying a hypothesis test to detect inconsistency between treatment effects estimated using direct and indirect evidence.

Loop specific approach: for 3 treatments, one test; for >3 treatments, one test per loop. Composite test for inconsistency: chi-square test for different evidence sources for same comparison (direct evidence and indirect evidence from multiple loops).

No statistically significant test result.

Indicates valid assumption.

Statistically significant test result.

Indicates invalid assumption.

- Multi-arm trials pre
usefulness of the r
- Results are influen
number of evidence
number of studies.

Method C7: The back transformation method

Back transformation method (method C7)

Using back calculation to estimate treatment effects based on indirect evidence (from the results of NMA and pair-wise MA) then comparing the direct and indirect estimates using a hypothesis test or visual comparison

- Multi-arm trials preclude the
- usefulness of the method.
- Visual interpretation is subjective.

- No statistically significant test result
  - Indicates valid assumption
- Treatment effects using direct and indirect evidence are similar (using visual interpretation)
  - Indicates valid assumption
- Statistically significant test result
  - Indicates invalid assumption
- Treatment effects using direct and indirect evidence differ (using visual interpretation)
  - Indicates invalid assumption

Using back calculation to estimate treatment effects based on indirect evidence (from the results of NMA and pair-wise MA) then comparing the direct and indirect estimates using a hypothesis test or visual comparison.

- Multi-arm trials preclude the
- usefulness of the method.
- Visual interpretation is subjective.

No statistically significant test result.

Treatment effects using direct and indirect evidence are similar (using visual interpretation).

Statistically significant test result.

Treatment effects using direct and indirect evidence differ (using visual interpretation).

Indicates valid assumption

Indicates invalid assumption

Method C8: Multidimensional scaling

Multidimensional scaling (method C8)

Calculating the inverse variance average of the treatment effects across trials for each treatment pairing (called observed dissimilarities) and using weighted least squares scaling to obtain the corresponding fitted dissimilarities.

- Observed and fitted dissimilarities are comparable.
  - Indicates valid assumption.
- Observed and fitted dissimilarities are discrepant.
  - Indicates invalid assumption.

- Multi-arm trials may preclude the usefulness of the method.

Calculating the inverse variance average of the treatment effects across trials for each treatment pairing (called observed dissimilarities) and using weighted least squares scaling to obtain the corresponding fitted dissimilarities.

Observed and fitted dissimilarities are comparable.

Indicates valid assumption.

Observed and fitted dissimilarities are discrepant.

Indicates invalid assumption.

- Multi-arm trials may preclu
the usefulness of the met

Method C9: The two-stage approach

Two-stage approach (method C9)

Applying a likelihood ratio test to detect overall inconsistency amongst evidence sources and visually assessing the magnitude of residuals.
Group trials by design and pair-wise meta-analyse.
Estimate NMA estimates by performing meta-regression on the direct estimates using the design matrix.

- Visual interpretation is subjective.

No statistically significant likelihood ratio test result.

No obvious outlying or large residuals.

Statistically significant likelihood ratio test result.

Outlying or large residuals identified.

Indicates valid assumption.

Indicates invalid assumption.

Two Stage approach Method 3/7

Applying a likelihood ratio test to detect overall inconsistency amongst evidence sources and visually assessing the magnitude of residuals.
Group trials by design and pair-wise meta-analyse.
Estimate NMA estimates by performing meta-regression on the direct estimates using the design matrix.

Visual interpretation is subjective.

No statistically significant likelihood ratio test result.

Indicates valid assumption.

No obvious outlying or large residuals.

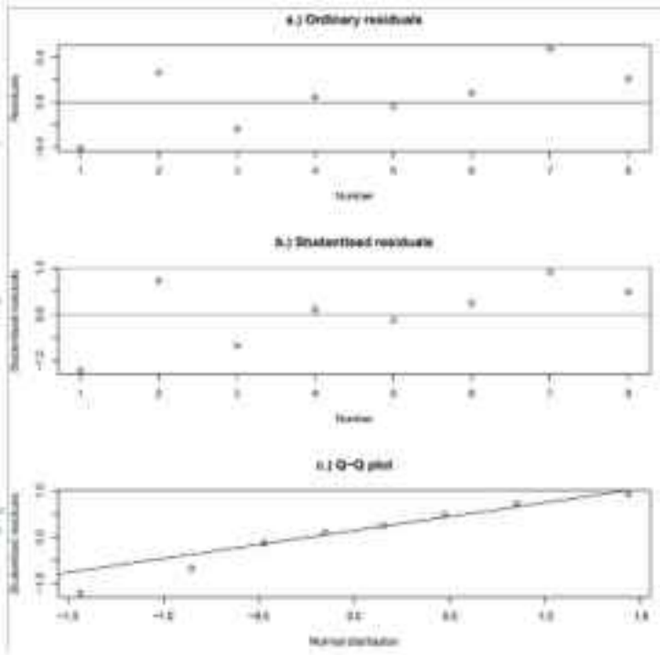
Statistically significant likelihood ratio test result.

Indicates invalid assumption.

Outlying or large residuals identified.

Two-stage approach (method C9)

There are no obviously large ordinary residuals



All studentised residuals fall within the interval (-196; 196) indicating good consistency

The Q-Q plot for the studentised residuals justifies the model assumption of normality

Design group	Treatments	Number of sites (number of patients)	Comparison	First stage log odds ratio (95% confidence interval)	Second stage log odds ratio (95% confidence interval)	Residuals	Studentised residuals
1	DHAPQ, AQ+AS	2 (332)	AQ+AS vs. DHAPQ	-1.46 (-2.28, -0.64)	-1.05 (-1.54, -0.57)	-0.41	-1.21
2	DHAPQ, AQ+AS, AL	5 (1080)	AQ+AS vs. DHAPQ	-0.79 (-1.65, 0.07)	-1.05 (-1.54, -0.57)	0.26	0.73
			AL vs. DHAPQ	-1.07 (-1.94, -0.20)	-0.83 (-1.34, -0.32)	-0.24	-0.67
3	DHAPQ, AQ+AS, CD+A	2 (461)	AQ+AS vs. DHAPQ	-1.01 (-1.92, -0.10)	-1.05 (-1.54, -0.57)	0.04	0.11
			CD+A vs. DHAPQ	-2.14 (-3.02, -1.26)	-2.10 (-2.74, -1.47)	-0.04	-0.12
4	DHAPQ, AL	4 (688)	AL vs. DHAPQ	-0.75 (-1.55,0.05)	-0.83 (-1.34, -0.32)	0.08	0.25
5	DHAPQ, AL, CD+A	4 (679)	AL vs. DHAPQ	-0.36 (-1.48, 0.76)	-0.83 (-1.34, -0.32)	0.47	0.93
			CD+A vs. DHAPQ	-1.90 (-2.94, -0.86)	-2.10 (-2.74, -1.47)	0.21	0.50

Likelihood ratio test
Test statistic= 2.69, df=5, P=0.75
Statistically significant inconsistency is not detected

The results do not indicate the presence of inconsistency.

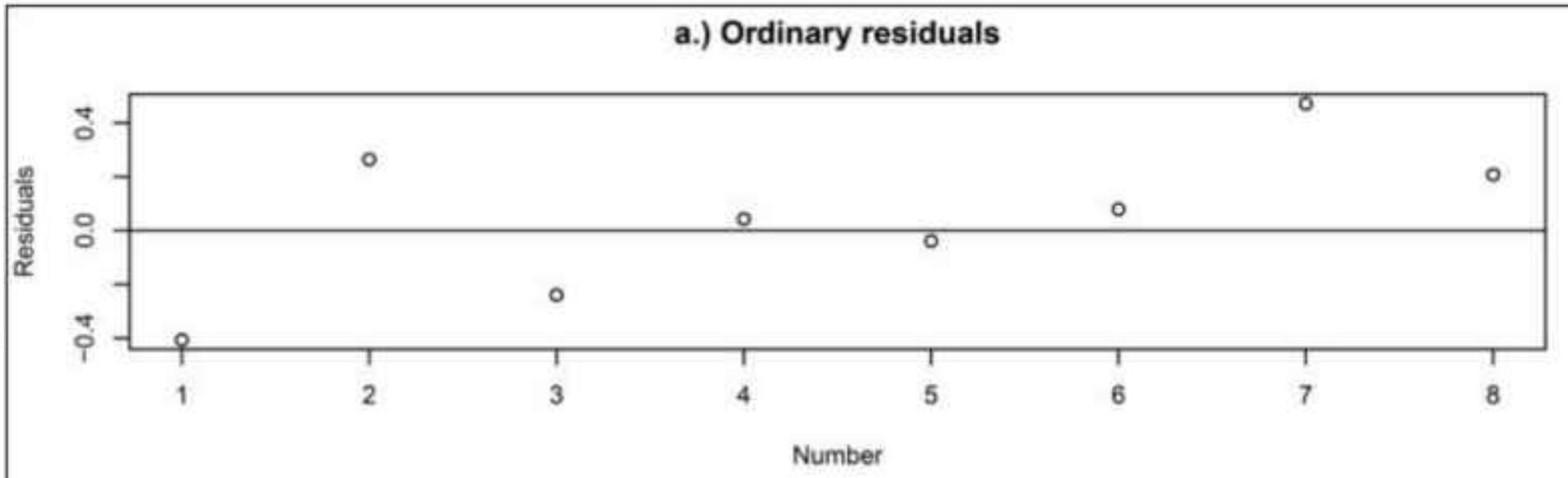
Design group	Treatments	Number of sites (number of patients)	Comparison	First stage log odds ratio (95% confidence interval)	Second stage log odds ratio (95% confidence interval)	Residuals	Studentised residuals
1	DHAPQ, AQ+AS.	2 (332)	AQ+AS vs. DHAPQ	-1.46 (-2.28, -0.64)	-1.05 (-1.54, -0.57)	-0.41	-1.21
2	DHAPQ, AQ+AS, AL.	5 (1690)	AQ+AS vs. DHAPQ	-0.79 (-1.65, 0.07)	-1.05 (-1.54, -0.57)	0.26	0.73
			AL vs. DHAPQ	-1.07 (-1.94, -0.20)	-0.83 (-1.34, -0.32)	-0.24	-0.67
3	DHAPQ, AQ+AS, CD+A.	2 (491)	AQ+AS vs. DHAPQ	-1.01 (-1.92, -0.10)	-1.05 (-1.54, -0.57)	0.04	0.11
			CD+A vs. DHAPQ	-2.14 (-3.02, -1.26)	-2.10 (-2.74, -1.47)	-0.04	-0.12
4	DHAPQ, AL.	4 (688)	AL vs. DHAPQ	-0.75 (-1.55,0.05)	-0.83 (-1.34, -0.32)	0.08	0.25
5	DHAPQ, AL, CD+A.	4 (679)	AL vs. DHAPQ	-0.36 (-1.48, 0.76)	-0.83 (-1.34, -0.32)	0.47	0.93
			CD+A vs. DHAPQ	-1.90 (-2.94, -0.86)	-2.10 (-2.74, -1.47)	0.21	0.50

Likelihood ratio test

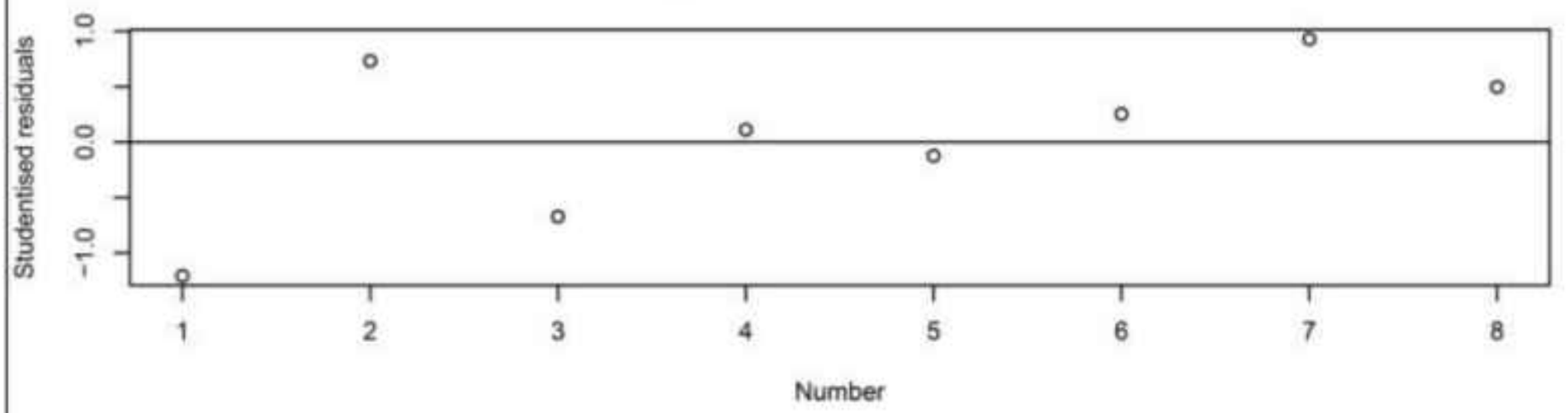
Test statistic= 2.69, df=5, P=0.75.

Statistically significant inconsistency is not detected.

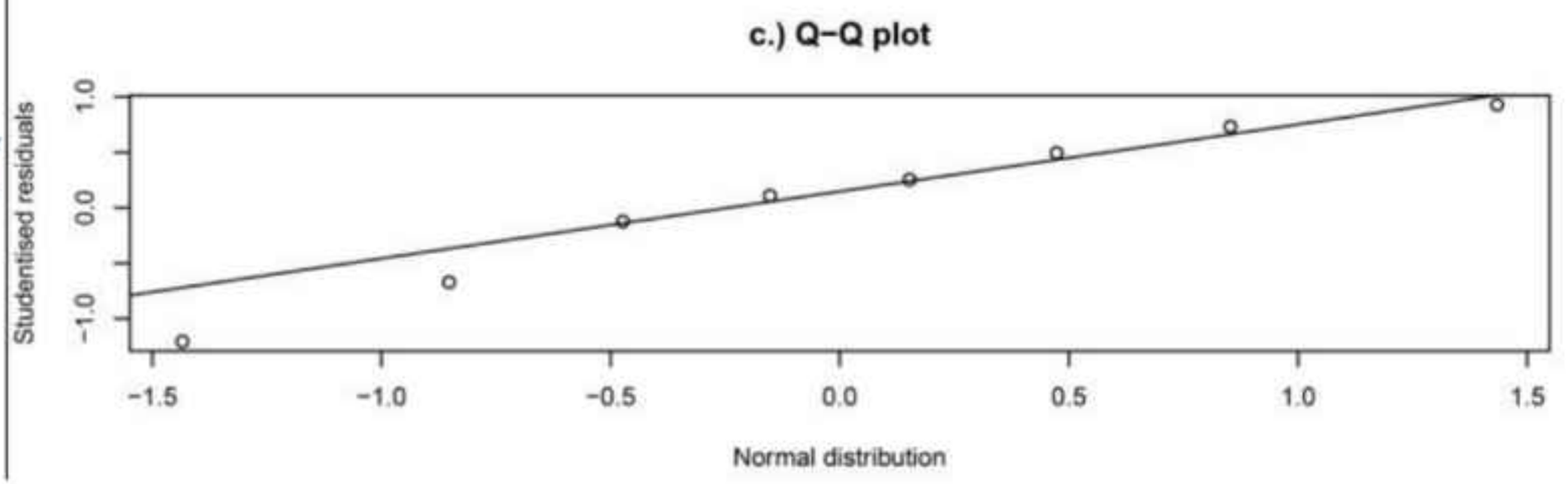
There are no obviously large ordinary residuals.



All studentised residuals fall within the interval (-1.96, 1.96) indicating local consistency.



The Q-Q plot for the studentised residuals justifies the model assumption of normality.



Method C10: The graph-theoretical method

Graph-theoretical method (method C10)

Applying graph theoretical methods to estimate a measure of residual inconsistency.

Q statistic measures the difference between the observed treatment effects (from trials) and the potential differences.
Potential differences are the treatment effects that guarantee consistency (calculated using the observed treatment effects and laws of electrical theory).
A measure of consistency given by calculating the Q statistic for each pairwise comparison in the NMA, summing them and subtracting the sum from Q statistic from the NMA.

- Subjective
- Currently, multi-arm trials preclude the usefulness of the method.
- Currently, fixed treatment effects only.

Small residual inconsistency.

Large residual inconsistency.

Indicates valid assumption.

Indicates invalid assumption.

Applying graph theoretical methods to estimate a measure of residual inconsistency.

Q statistic measures the difference between the observed treatment effects (from trials) and the potential differences.
Potential differences are the treatment effects that guarantee consistency (calculated using the observed treatment effects and laws of electrical theory).
A measure of consistency given by calculating the Q statistic for each pairwise comparison in the NMA, summing them and subtracting the sum from Q statistic from the NMA.

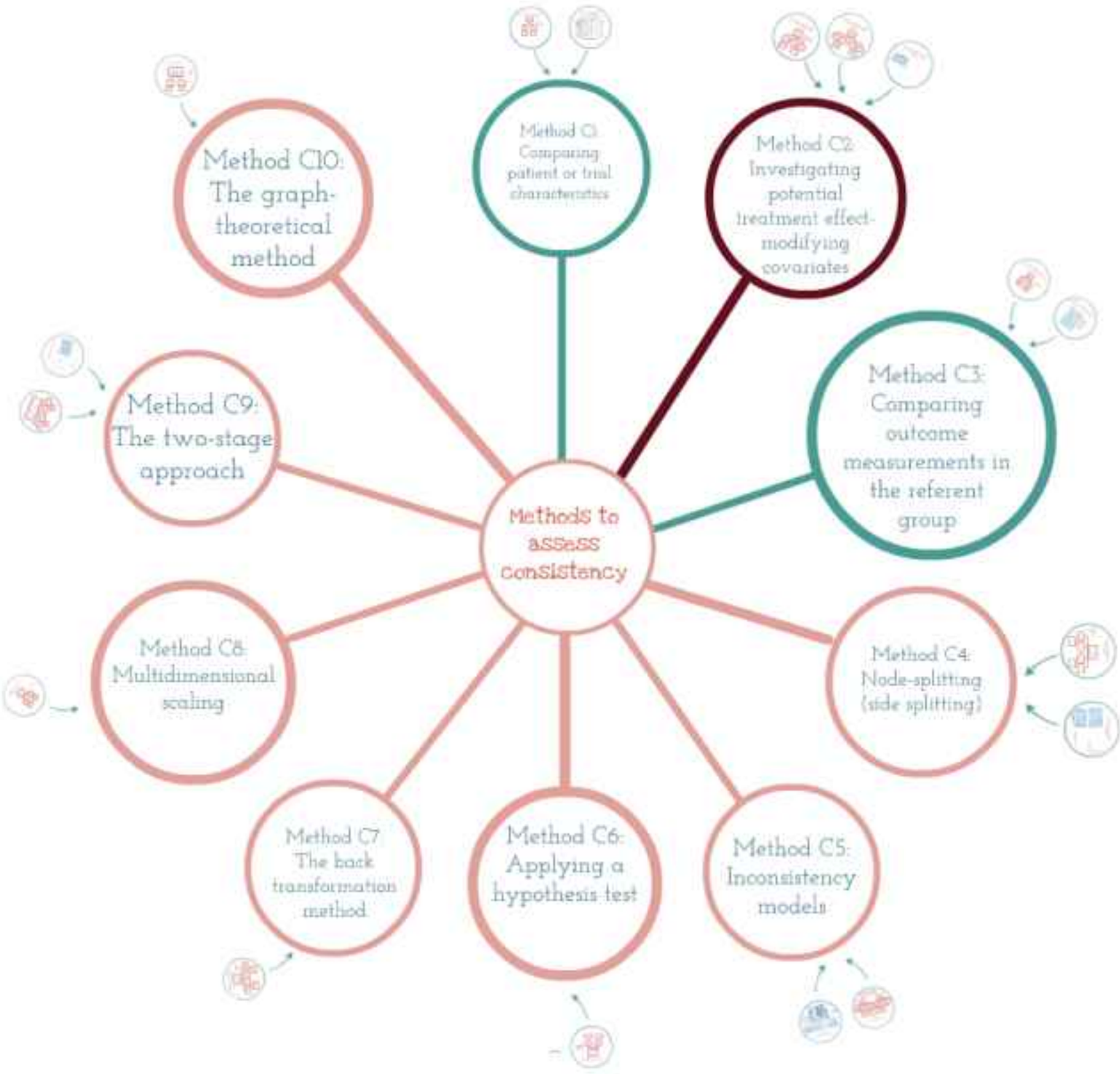
Small residual inconsistency.

Indicates valid assumption.

Large residual inconsistency.

Indicates invalid assumption.

- Subjective
- Currently, multi-arm trials preclude the usefulness of the method.
- Currently, fixed treatment effects only.


Methods to assess consistency
Method C1: Comparing patient or trial characteristics
Method C2: Investigating potential treatment effect-modifying covariates
Method C3: Comparing outcome measurements in the referent group
Method C4: Node-splitting (side splitting)
Method C5: Inconsistency models
Method C6: Applying a hypothesis test
Method C7: The back transformation method
Method C8: Multidimensional scaling
Method C9: The two-stage approach
Method C10: The graph-theoretical method

Things to note....

All results presented are from models
with random treatment effects.
that account for correlation in multi-arm trials.
that assumes homogeneous between trial variances.

Results were estimated using a bayesian approach (Winbugs) except for the 2 stage approach.

Review's conclusions

'Presently, we advocate applying existing assessment methods collectively to gain the best understanding possible regarding whether the assumption is reasonable.'

'Of course, the methods applied must be appropriate to the dataset, for example, the back transformation method is unsuitable when multi-arm trials contribute.'

- How should a review author proceed if the assessment methods indicate that the consistency assumption may be unrealistic?
- What if, despite the review author's best efforts, the cause of inconsistency is untraceable?
- Perhaps, if the consistency assumption seems unrealistic, NMA should be avoided; inconsistency could be accounted for using inconsistency modelling; or explained using meta-regression techniques?
- Moreover, if one or more assumption does not seem appropriate, how unreliable are the results?
- Is user -friendly software needed?

Review's conclusions

'Presently, we advocate applying existing assessment methods collectively to gain the best understanding possible regarding whether the assumption is reasonable.'

'Of course, the methods applied must be appropriate to the dataset, for example, the back transformation method is unsuitable when multi-arm trials contribute.'

aset, for example, the back
tribute.'

- How should a review author proceed if the assessment methods indicate that the consistency assumption may be unrealistic?
- What if, despite the review author's best efforts, the cause of inconsistency is untraceable?
- Perhaps, if the consistency assumption seems unrealistic, NMA should be avoided; inconsistency could be accounted for using inconsistency modelling; or explained using meta-regression techniques?
- Moreover, if one or more assumption does not seem appropriate, how unreliable are the results?
- Is user -friendly software needed?