A protocol for constructing a tool to assess clinical heterogeneity in meta-analyses, assessment of interrater variability, and a pilot study of the association between clinical and statistical heterogeneity

Marija Barbateskovic[1,4], Thijs M. Koster[3], Frederik Keus[3], Christian Gluud[1], Morten H. Møller[2,4], Iwan C.C. van der Horst[3], Anders Perner[2,4], Jørn Wetterslev[1,4].

1. The Copenhagen Trial Unit (CTU), Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
2. Department of Intensive Care 4131, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark
3. Department of Critical Care, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
4. Centre for Research in Intensive Care (CRIC), Copenhagen, Denmark

Objectives

1) To construct a tool to assess clinical heterogeneity in meta-analyses.

2) To assess the agreement between independent use of a new tool for assessing clinical heterogeneity.

3) To explore the association between clinical and statistical heterogeneity in 40 Cochrane Review meta-analyses of interventions in assorted medical fields.

Introduction

Systematic reviews and meta-analyses of randomised clinical trials assessing relevant clinical interventions are used to answer questions of whether the interventions benefit or harm patients, and they frequently inform guidelines and practice protocols [1]. Several methods exist for estimating statistical heterogeneity (statistical between trial variance) when performing meta-analyses, but few methods are available for estimating overall clinical heterogeneity (e.g. different populations, different doses, different duration of applied intervention, etc.) [2]. Nonetheless, it is recommended always to consider clinical heterogeneity when performing systematic reviews and meta-analyses [3]. Clinical heterogeneity may hamper the interpretation of results from meta-analyses in systematic reviews and statistical heterogeneity affect the uncertainty of the estimates of the intervention effects [3]. A priori assumptions of clinical heterogeneity may influence the choice of the primary statistical model to be applied [3], e.g. fixed-effect or random-effects meta-analyses [4], and these assumptions may or may not be confirmed when the statistical heterogeneity, the between trial variance, of the included trials has been estimated.

We have established a database containing nearly all systematic reviews and meta-analyses of interventions used in the Intensive Care Unit (ICU) including patients from ICUs. This database has been established in The Critical Care Research Unit at University Medical Center, Groningen, [5] and provided us with experience on meta-analyses of interventions

used in the ICU and the seemingly high degree of clinical heterogeneity within these. We therefore want to explore the clinical and statistical heterogeneity in meta-analyses in general, in ICU meta-analyses, and whether heterogeneity within Cochrane ICU meta-analyses contrast with heterogeneity within Cochrane meta-analyses of interventions focusing on clinical scenarios outside the ICU.

We have chosen to investigate meta-analyses of interventions used in the ICU in our unique collection of meta-analyses as the burden for patients and relatives of admission to the ICU is tremendous [6], mortality among ICU patients is high, and the survivors may have a high risk of impaired quality of life [7]. Still, it is expected that in the coming years more patients will need intensive care due to an aging population and advances in medical and surgical care. Intensive care is costly and the increasing need for intensive care will enlarge the demands on the health care system [8].

Due to suspected lack of power to assess the association between clinical heterogeneity and statistical heterogeneity in 40 meta-analyses from Cochrane Reviews, the study should only be considered hypothesis generating on the possible association between clinical and statistical heterogeneity.

Heterogeneity in meta-analyses from systematic reviews

Meta-analyses in systematic reviews have become one of the most widely used methods to quantify the effects of clinical interventions. They are recognized as the best available evidence for decisions about health-care management and policy [1]. Nonetheless, it appears that health-care professionals and policy makers infrequently use systematic reviews to guide decision making [1, 11, 12]. One proposed reason for this, is the missing assessment of sources of clinical heterogeneity leading to inconclusive and non-specific results [13].

Several possible sources of heterogeneity exist among trials included in systematic reviews and meta-analyses. Clinical heterogeneity is characterised by variability in settings, participants, the interventions and comparators characteristics, the use of co-interventions, and the types and timing of outcome measurements. Methodological heterogeneity or risk of bias is characterised by variability in trial design and quality in seven distinct domains [3],

and statistical heterogeneity is characterised by variability in summary treatment effects between trials. The presence and magnitude of statistical heterogeneity is associated with risk of bias [14, 15, 16, 17, 18] and/or may be associated with clinical sources of heterogeneity [14,15], arise from other unknown or unrecorded trial characteristics (residual variation), or simply random errors ('play of chance') due to lack of data and trials. In the context of systematic reviews, clinical heterogeneity can be defined as the differences in clinically related trial characteristics which can give rise to variations in pooled treatment effects estimates not covered by the bias assessment of the included trials [13,14, 15].

In contrast to guidance on the assessment and investigation of methodological and statistical heterogeneity [2], little attention has been given to the common clinical heterogeneity in systematic reviews [16,17, 18] except for planned subgroup analyses and meta-regression analyses which may address differences in trial characteristics in univariate or multivariate analyses in an unstandardized way. Further, subgroup analyses and meta-regression analyses will not be able to assess overall summary clinical heterogeneity. It therefore seems crucial to increase our understanding of clinical heterogeneity to interpret systematic reviews and it seems important to investigate if clinical heterogeneity is associated with statistical heterogeneity. Furthermore, as methodological heterogeneity in terms of bias does not include several differences of the use of the interventions such as: different intensities of the experimental intervention (doses, length of intervention, etc.); timing of interventions; use or no use of co-interventions; use or no use of different control interventions; definition and timing of outcome measurement (length of follow-up), it may be important to quantify such heterogeneity as it is not characterised in the usual methodological or bias assessment.

In this protocol we will define and subgroup clinical heterogeneity in four domains: setting heterogeneity, population heterogeneity, intervention heterogeneity, and outcome heterogeneity, which, besides variation in included trial populations, include methodological differences of the trial characteristics not necessarily covered by the bias domains according to the Cochrane Handbook [3]. Further, we will:
1) develop and present a new scoring system for assessing the degree of clinical heterogeneity;

2) present a plan for assessing the interrater agreement between independent users of this scoring system;
3) present a plan for assessing the association between clinical heterogeneity and statistical heterogeneity in 40 meta-analyses from Cochrane reviews of interventions in assorted medical fields, 20 ICU meta-analyses and 20 non-ICU meta-analyses [10].

Development of a clinical heterogeneity in meta-analysis score and association between clinical and statistical heterogeneity in meta-analyses

During the protocol phase, we aim to construct a clinical heterogeneity for a meta-analysis score (CHIMS) in systematic reviews during a pilot phase based on the works of Gagnier et al in 2011 [19] and 2013 [20]. The pilot phase will involve scoring of 3 meta-analyses, adjusting of incomprehensible phrasing of items and development of a manual for using CHIMS, another scoring of 5 meta-analyses followed by an adjustment for categorizing CHIMS into low, moderate or unclear, and high clinical heterogeneity. After the protocol phase, the CHIMS will be validated in a second project for inter-observer variability by two independent evaluators involved in the development of CHIMS, thereafter by two independent evaluators not involved in the development of CHIMS, and finally by 20 pairs of review authors scoring a meta-analysis of the primary, dichotomous outcome of their own systematic review.

In the third project, we aim to investigate the association between clinical heterogeneity and statistical heterogeneity in critically ill patients within 20 ICU Cochrane review meta-analyses 20 non-ICU Cochrane review meta-analyses.

First project

Constructing and developing the CHIMS during a pilot phase of assessing 3 meta-analyses with subsequent adjustment of the CHIMS, writing a manual for guiding the use of the CHIMS, and hereafter assessing 5 meta-analysis for categorizing the summary CHIMS into low, moderate or unclear, and high clinical heterogeneity. This project has resulted in the tool presented in Table 1 at page 8, the categories of low, moderate or unclear, and high heterogeneity presented in 'Data-synthesis and statistical analysis' at page 11, and a manual for the use of the CHIMS presented in the Appendix.

Second project

1. The interrater agreement of CHIMS assessed by two independent users involved in the development of the CHIMS and the CHIMS manual in 40 Cochrane meta-analyses, 20 ICU meta-analyses and 20 non-ICU meta-analyses. We will only investigate the meta-analysis on the primary dichotomous outcome from each systematic review. After individual and independent scoring of CHIMS the evaluators will pairwise agree upon a common CHIMS which will be used in the second project part 5. And in the third project.
2. The interrater agreement of CHIMS assessed by two independent users not involved in the development of the CHIMS and the CHIMS manual in 40 Cochrane meta-analyses, 20 ICU meta-analyses and 20 non-ICU meta-analyses. After individual and independent scoring of CHIMS the evaluators will pairwise agree upon a common CHIMS which will be used in the second project part 5. And in the third project.
3. The interrater agreement in 20 systematic reviews within 20 pairs of review authors scoring a meta-analysis of the primary, dichotomous outcome of their own systematic review. After individual and independent scoring of CHIMS the evaluators will pairwise agree upon a common CHIMS which will be used in the second project part 5. And in the third project.
4. The interrater agreements from 1), 2), and 3) with 95% confidence intervals, assessed by weighted Kappa for agreement between two users of the CHIMS, with unweighted items scores, of the clinical heterogeneity in 40 meta-analyses with low CHIMS (score 0-10), moderate or unclear CHIMS (score 11-18), and high CHIMS (score 19-22) will be calculated.
5. We will in a supplementary exploratory analysis estimate the agreement between low (0-30%), moderate (30-60%), and high (60-100%) statistical heterogeneity (between trial heterogeneity rather than sampling error I-square (I^2) [22] or D-square (D^2) [23]) and low(score 0-10), moderate or unclear (score 11-18), and high CHIMS (score 19-22) using the CHIMS agreed upon within each pair of evaluators.

Third project

1. Primary analysis: association between the statistical heterogeneity (I^2, D^2, and Tau^2) in random-effects meta-analyses of primary outcomes [4] and the CHIMS agreed upon by the evaluators of the CHIMS investigated by log-linear regression adjusted for risk of bias and control event proportion. We will only investigate the meta-analysis on the primary dichotomous outcome from each systematic review.
2. Secondary analysis: association between the variance of the pooled intervention effect in a random-effects meta-analysis [4] and the CHIMS and the CHIMS agreed upon by the evaluators of the CHIMS investigated by log-linear regression adjusted for risk of bias and control event proportion.
3. We will investigate the association between the CHIMS agreed upon by the evaluators of the CHIMS and statistical heterogeneity within meta-analyses with low risk of bias and within meta-analyses with high risk of bias.
4. We will investigate the association between the CHIMS agreed upon by the evaluators of the CHIMS and statistical heterogeneity within meta-analyses showing statistically significant effects on primary outcomes and within reviews and meta-analyses not showing statistically significant effects on primary outcomes.

Material and methods

This project will be led from Copenhagen Trial Unit, within The Centre for Research in Intensive Care (CRIC) in cooperation with Research Unit at the Department of Critical Care of the University Medical Center Groningen in The Netherlands. Below we describe the general methods which will be used in the planned project.

Criteria for considering meta-analyses for inclusion

Based on 60 Cochrane systematic reviews of interventions for ICU patients (included in the database established by Thijs Koster and Frederic Keus [5], containing systematic reviews and meta-analyses of interventions used in the ICU including patients from ICUs established in The Critical Care Research Unit at University Medical Center Groningen [5]) we will select 20 Cochrane systematic reviews not already selected for developing the CHIMS. Another 20 Cochrane systematic reviews of interventions used outside the ICU will be selected to cover a wide range of non-ICU interventions from different Cochrane review groups published in The Cochrane Library. We will choose these meta-analyses from the last edition and go

backwards from date of publication, if not too late, thereby selecting the most updated meta-analyses. We will only select meta-analyses assessing a primary, dichotomous outcome with 3 or more randomised clinical trials included.

Assessment of risk of bias in meta-analysis and trials

We will use the bias risk assessment of trials and GRADE assessment [16] of the evidence for the effect estimates performed by the authors of the included Cochrane reviews. We will classify these meta-analyses as overall high risk of bias [16] if the meta-analysis includes one or more trials with overall high risk of bias [9, 17] and as low risk of bias if all included trials are overall low risk of bias.

First project: Protocol and pilot phase for development of the Clinical Heterogeneity in Meta-analysis Score (CHIMS)

During a pilot phase Marija Barbateskovic and Thijs Koster adjusted the phrasing and criteria for scoring 0, 1, and 2 points within the CHIMS items (see table 3), developed a manual (see Appendix) for conducting CHIMS, and proposed a categorization of CHIMS into low, moderate or unclear, and high clinical heterogeneity (see 'Data-synthesis and statistical analysis' second section at page 11).

Assessment of clinical heterogeneity

CHIMS will be measured on a scale including the following 4 domains with overall 11 selected items covering the most important domains and items describing clinical heterogeneity:

1. **Domain: Population heterogeneity (4)**: disease severity, comorbidities, age, and sex.
2. **Domain: Setting heterogeneity (1 item)**: period where the trial was performed/reported whether it was performed in a developed (D) or in a developing country (DC), or e.g. performed in a high dependence unit compared to ICU unit or a high dependence unit compared to standard unit, etc.
3. **Domain: Intervention heterogeneity (4 items)**: Intervention intensity (dose, frequency, duration, device, cut-off values), timing of intervention (number of times per time unit, continuous), heterogeneity of control-interventions in included trials, distribution of co-interventions in randomised groups.

4. **Domain: Outcome heterogeneity (2 items)**: Definition of outcome, timing of outcome assessment.

Overall CHIMS score: The 11 items in an unweighted CHIMS will be scored 0, 1, or 2 according to low=0, moderate or unclear=1, or high clinical heterogeneity=2, with a total range of 0 to 22 with equal weight assigned to each item:

Table 1

Domains	Items	Score	Explanation of score for extreme differences between trials in a meta-analysis
Population heterogeneity	**1. Age**	0	Mean/median age ≤10 years difference
		1	11-20 years difference in mean/median age
		2	More than > 20 years difference in mean/median age
	2. Sex	0	% women ≤20 % absolute difference between trials
		1	21-30% absolute difference of % women between trials
		2	More than > 30% absolute difference between trials
	3. Participant inclusion criteria and baseline disease severity	0	RCTs include patients that are equally ill or the difference in risk or score for disease severity of patients ≤20%
		1	Condition/patient population differs slightly with 50% or more overlap of types of participants and/or the difference in risk or score for disease severity of patients is 21-30%
		2	Condition/patient population differs considerable and/or the difference in risk or score for disease severity of patients >30% Use relative difference when inclusion criteria are assessed (disease severity scores).
	4. Co-morbidities	0	Difference in frequency of important comorbidities ≤20% or no co-morbidities are reported in the RCTs and differences in co-morbidities are assumed absent
		1	Slight differences in important co-morbidities, between 21 and 30%, or no co-morbidities are reported in the RCTs, but differences in co-morbidities are assumed
		2	Differences in frequency of important comorbidities > 30% or highly likely variations in co-morbidities Use absolute difference when comparing important co-morbidities.

Setting heterogeneity	**5. Years reported (A), performed in developed (D) vs developing country (DC) (B), unit type (C)**	0 1 2	No differences: A) years reported differ <15, B) No D vs DC, treating units similar, OR C) slight variations in the unit or facility type and there is low risk of affecting other fields of heterogeneity Slight variation: A) years reported differ ≥15, OR B) D vs Dc, OR treating units not similar, OR C) if there are slight variations in the unit or facility type but there is risk of affecting other fields of heterogeneity (at least one of A-C involved) Considerable variation: A) years reported differ ≥15, AND B) D vs. Dc, AND C) treating units not similar (all of A-C involved), AND/OR if the RCTs in the opinion of the assessor differs markedly in setting heterogeneity
Intervention heterogeneity	**6. Intensity, strengths, or duration of intervention**	0 1 2	Little variation: Differences in dose, strengths, devices, cut-offs, or duration of interventions ≤20% Slight variation: 21-30% differences in dose, strengths, devices, cut-offs, or duration intervention, or if dose, strength, cut-offs or duration of intervention cannot be assessed from the information in the RCTs Considerable variation: if different types of interventions are used, or different doses, strengths, devices, cut-offs, or duration of intervention >30% Use relative differences when assessing intensity, strengths, duration.
	7. Timing	0 1 2	Criteria for starting the intervention are similar, or relative differences of timing of intervention differs ≤20% Criteria for starting the intervention differ slightly, or the relative timing difference is 21-30% Criteria for starting the intervention differ, or relative timing difference exceeds >30%
	8. Control intervention	0 1 2	All control interventions are the same Control interventions include placebo AND no intervention, assess as item 6 if an active intervention is used Including trials with different active control interventions OR trials with active and placebo/no intervention

	9. Co-interventions	0 1 2 Use absolute difference when assessing co-interventions.	No apparent differences in co-interventions, OR standard care is not described or assumed to be the same, OR equally applied in groups, or different co-interventions are used but the effect of the co-intervention is assumed to be small Slight variation in co-interventions or the same co-interventions are used with slight variation (<30 %) Considerable differences if it is assumed that the co-intervention is not usual care, or differences in use of co-interventions >30 %
Outcome heterogeneity	**Definition of the outcome in the meta-analysis**	0 1 2	Same definition of outcome Slight variations in definition of outcome Considerable variations in definition of outcome
	Timing of outcome measurement	0 1 2	Less than one month between follow-up of outcome More than one but less than or equal to 3 months between follow-ups More than 3 months between follow-up of outcome

Explanation for the use of the clinical heterogeneity in meta-analysis score (CHIMS)

To guide evaluators of CHIMS we have provided somewhat arbitrary thresholds for the scores 0, 1, and 2 and a manual for using the CHIMS (see Appendix) which should help getting higher agreements between independent evaluators. However, we hope everyone can agree that e.g. more than 30% relative difference between different trials dose of a drug intervention or e.g. that more than 30% relative difference in risk of severe disease (or severity score) are substantial differences. However, difference between 20% and 30% are probably not substantial but may influence results, and less than 20% is probably less important.

If difference between trials for a specific item is impossible to detect or reject, we suggest that the meta-analysis score 1 on the given item corresponding to unclear.

We realise that the chance of meeting a higher score might be associated with the number of trials included in a meta-analysis, however, statistical heterogeneity may also increase

with more trials, not necessarily, but there is a trend that way, however, a ceiling effect for statistical heterogeneity also seems likely after 20 to 25 included trials [21].

Data-synthesis and statistical analysis

The interrater agreement of CHIMS assessed by two independent evaluators in 40 meta-analyses (20 ICU meta-analysis and 20 non-ICU meta-analysis), analysed by linear regression and weighted Kappa for agreement between two evaluators, with 95% confidence intervals, of the CHIMS, with unweighted items scores, of the clinical heterogeneity in 40 meta-analyses with clinical heterogeneity low (score 0-10), moderate or unclear (score 11-18), and high (score 19-22) will be calculated:

Second project

1) We will analyse the interrater agreement of CHIMS, assessed by two independent users involved in the development of the CHIMS and the CHIMS manual, in 40 meta-analyses, 20 ICU meta-analyses and 20 non-ICU meta-analyses, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression [22]. We will investigate the interaction between interrater agreement or ICC and whether the meta-analyses are ICU or non-ICU meta-analyses.
2) We will analyse the interrater agreement of CHIMS, assessed by two independent users not involved in the development of the CHIMS and the CHIMS manual, in 40 meta-analyses, 20 ICU meta-analyses and 20 non-ICU meta-analyses, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression [22]. We will investigate the interaction between interrater agreement or ICC and whether the meta-analyses are ICU or non-ICU meta-analyses [22].
3) We will analyse the interrater agreement, in 20 systematic reviews within 20 pairs of review authors scoring a meta-analysis of the primary, dichotomous outcome of their own systematic review, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression [22]. We will investigate the interaction between interrater agreement or ICC and whether the meta-analyses are ICU or non-ICU meta-analyses.

4) We will analyse the interrater agreements from 1), 2), and 3) with 95% confidence intervals, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression [22] for agreement between two users of the CHIMS score, with unweighted items scores, of the clinical heterogeneity in 40 meta-analyses with low CHIMS (score 0-10), moderate or unclear CHIMS (score 11-18), and high CHIMS (score 19-22). We will investigate the interrater agreement in the ICU meta-analyses and the non-ICU meta-analyses.
5) Secondary analysis: We will in a supplementary, exploratory, analysis estimate the agreement, with weighted Kappa, between low, moderate, and high statistical heterogeneity (between trial heterogeneity rather than sampling error I^2 [23] or D^2 [24]) and low, moderate or unclear, and high clinical heterogeneity.

Third project

1) Primary analysis: analysis of association between the statistical heterogeneity (I-square, D-square, and Tau-square) in random-effects meta-analyses of primary outcomes [4] and the CHIMS score of clinical heterogeneity adjusted for risk of bias, ICU or non-ICU MA, and control event proportion. Interactions between the possible association and the adjusting co-variates will be explored. We will only investigate the MA on the primary, dichotomous outcome from each systematic review.

2) Secondary analysis: association between the total variance of the pooled intervention effect in a random-effects meta-analysis [4] and the CHIMS score of clinical heterogeneity investigated by log-linear regression adjusted for risk of bias, ICU or non-ICU MA, and control event proportion. Interactions between the possible association and the adjusting co-variates will be explored. We will only investigate the MA on the primary, dichotomous outcome from each systematic review.
3) We will investigate the association between clinical heterogeneity and statistical heterogeneity within meta-analyses with low risk of bias and within meta-analyses with high risk of bias.
4) We will investigate the association between clinical heterogeneity and statistical heterogeneity within meta-analyses showing statistically significant

effects on primary outcomes and within reviews and meta-analyses not showing statistically significant effects on primary outcomes.

If possible, we will investigate the agreement between the categorised clinical heterogeneity and statistical heterogeneity within meta-analyses showing statistically significant effects (P<0.05 or with breakthrough of the trial sequential monitoring boundary for benefit or harm in a Trial Sequential Analysis (TSA) [25, 26,27]) on primary outcomes within reviews and meta-analyses not showing statistically significant effects (P>=0.05 or with no breakthrough of the TSA boundary for benefit or harm [25, 26,27]) on primary outcomes.

Investigating the association between clinical heterogeneity and between-trial statistical heterogeneity by using regression analyses of statistical heterogeneity versus clinical heterogeneity

We will explore a model:

$$\text{Log}\,(H_i) = \alpha_0 + \alpha_1 \cdot CH_i + \alpha_2 \cdot \text{Bias}_i + \alpha_3 \cdot CEP_i + \varepsilon_i$$

We will use log-linear regression to explore a possible association between statistical heterogeneity (H_i) and clinical heterogeneity (CH_i) where H_i in the i-*th* meta-analysis are depicted by I^2 (or D^2) which describes between trial variance relative to the total variance in a meta-analysis or τ^2 describing the absolute between trial variance. Where the α_j (j=0,…,3) are the regression coefficients, CH_i is the score of clinical heterogeneity in the *i-th* meta-analysis (0-20), Bias_i is the GRADE classification (0 for low risk of bias and 1 for high risk of bias) of the *i-th* meta-analysis according to bias risk of the meta-analysis, CEP_i is the unweighted control event proportion, and ε_i is the residual.

We will explore a possible association between the log standard error of mean of the pooled intervention effects (LogSEM(RR_i)= Log SEM(RR)) of the *i-th* meta-analysis *i*=1,….,60) and the

scores of clinical heterogeneity adjusted for association with bias (GRADE evaluation of meta-analytic bias, 0 being a meta-analysis of exclusively trials with overall low risk of bias an 1 a meta-analysis with at least one trial being overall high risk of bias), control event proportion (CEP_i), and information size (N_i) in the included meta-analyses and/or variance component analysis. The meta-variance-regression model of the association between LogSEM(RR_i) and the score of clinical heterogeneity CH_i in the *i*-th meta-analysis (0-15) will be investigated using univariate and multivariate Log-linear regression, the ultimate model being:

$$\text{Log SEM}(RR_i) = \alpha_0 + \alpha_1 \cdot CH_i + \alpha_2 \cdot Bias_i + \alpha_3 \cdot CEP_i + \alpha_4 \cdot N_i + \varepsilon_i$$

Where the α_j (j=0,…,3) are the regression coefficients, CH_i is the score of clinical heterogeneity in the *i*-th meta-analysis (0-20), $Bias_i$ is the GRADE classification (0 for low risk of bias and 1 for high risk of bias) of the *i-th* meta-analysis according to bias risk of the meta-analysis, CEP_i is the unweighted control event proportion, and N_i the actual accumulated information size in the *i-th* meta-analysis. ε_i is the residual.

We will explore the models fit in distribution of residual as well as the distributions of H_i , Log (H_i), SEM(RR_i), and the transformation to Log SEM(RR_i) to detect serious deviations from the normal distribution in which case we may abandon the parametric regression analyses and apply non-parametric regression analyses or other relevant transformations of the H_i and SEM(RR_i).

Possible associations between statistical heterogeneity and the 3 separate domains of CHIMS will be explored.

Possible associations in subgroup analyses according to risk of bias will be explored.

If possible, we will investigate the association between clinical heterogeneity and statistical heterogeneity within subgroups of meta-analyses showing conventional statistically significant effects (P<0.05) or with breakthrough of the TSA boundary for benefit or harm [25, 26,27]) on primary outcomes and within reviews and meta-analyses not showing

statistically significant effects (P >=0.05 or with no breakthrough of the TSA boundary for benefit or harm [25, 26,27]) on primary outcomes.

95% CIs will be calculated for Kappa and regression coefficients. P <0.05 will be considered statistically significant.

Appendix:

Manual for assessing Clinical Heterogeneity using CHIMS

CHIMS has been developed to detect and quantify clinical heterogeneity in meta-analyses. When difference has been identified between two trials for an item resulting in the score 2, it is not necessary to investigate the remaining RCTs, but one may move on to the next item.

In general, if differences between trials for a specific item are impossible to detect or reject, we suggest that the meta-analysis score 1 on the given item.

The percentage differences given below are relative unless otherwise stated

1.Domain Population heterogeneity

This domain is defined by multiple subdomains namely;

1. Age
2. Sex
3. Baseline disease severity
4. Co-morbidities

1. Age

Assess the difference in mean age between trials. If mean age is only given for each of the groups in a single RCT, calculate mean age of the included populations.

- Score 0: Mean/median age ≤ 10 years difference
- Score 1: 10-20 years difference in mean/median age
- Score 2: More than 20 difference in mean/median age

2. Sex

Assess the difference in sex bet ween trials. If sex is only reported for each of the groups in a single RCT, then calculate percentage of total males/females.

- Score 0: % women ≤ 20 % absolute difference between trials
- Score 1: 20% <% women <30 % absolute difference between trials
- Score 2: % women ≥ 30% absolute difference between trials

Example 1. RCT 1 included 20% females and RCT 2 included 45% females.
The absolute difference is 25%, therefore score a 1.

3. Baseline disease severity or differences in participants inclusion criteria

This subdomain assesses possible differences in patient diseases and the severity of these diseases.

Signalling question 1: Do the RCTs include the same type of participants or do the RCTs have similar inclusion criteria?

- Score 0: If the inclusion criteria of the RCTs describe the same types of participants.
- Score 1: If inclusion criteria differ slightly with 50% or more overlap of types of participants
- Score 2: If the RCTs include different types of patients.

Signalling question 2: How do the patients with the same inclusion criteria in the RCTs compare to each other? Are the patients' conditions similar? Are the patients in the RCTs equally 'ill'?

- Score 0: If the RCTs include patients that are equally ill.
- Score 1: If the RCTs include slightly different patients based on their illness, between 20 and 30 % difference in score for disease severity.
- Score 2: If the difference exceeds 30%.

Example 1. MA on the use of desmopressin for nocturia.
RCT 1 Includes men with voiding > 2/night.
RCT 2 men aged 40 to 65 years with LUTS, IPSS ≥ 13, persistent nocturia (≥ 2 episodes/night), nocturia index score ≥ 1 despite use of alpha-blocker treatment for ≥ 8 weeks, and nocturnal polyuria defined as nocturnal polyuria index > 33%.
The inclusion criteria differ between the two studies. The criteria of RCT 1 are less strict than in RCT 2. The participants included in the studies will assumedly be different. This item scores 1 point.

Example 2. MA on the use of dopamine for blood pressure regulation in ICU patients.
RCT 1 includes all patient admitted to the ICU with septic shock. Mean APACHE II (ICU

mortality score) score of 17 points.

RCT 2 includes all patient admitted to the ICU with septic shock. Mean APACHE II score of 15 points.

The patient inclusion criteria is the same. However, the disease severity differs. The difference in severity in this MA is 12%, thus this MA will score 1 point.

Example 3. MA on the use of honey as intervention in wound treatment.

RCT 1 includes patients with burn injury

RCT 2 includes patients post caesarean section or hysterectomy.

These patient categories differ. Thus, this MA scores 2 points on this item.

4. Co-morbidities

Co-morbidities are defined as the characteristics on diseases of the patients besides the inclusion criteria of the RCT that is included in the MA.

Signalling question 1: Are co-morbidities reported in the RCTs?

- Score 0: If no co-morbidities are reported in the RCTs and it is not assumable there are differences in co-morbidities.
- Score 1: If no co-morbidities are reported in the RCTs, but it is assumable that there are differences in co-morbidities.

Signalling question 2: If co-morbidities are reported, are the co-morbidities equally presented in the trials?

- Score 0: If there are little differences in clinical important comorbidities, less than 20%.
- Score 1: If there are slight differences, between 20 and 30%.
- Score 2: If there are important differences, more than 30%.

Example 1. MA on the use of dopamine for blood pressure regulation in ICU patients.

RCT 1 includes trauma patients and no reporting of co-morbidities

RCT 2 includes post cardiac surgery patients and no reporting of co-morbidities.

It is assumable that there are differences in co-morbidity between the trials, for example

renal function pre-trial admission probably differs between the two groups. However, it is not stated, thus score a 1.

Example 2. MA on the use of antibiotic prophylaxis in mechanical ventilated patients.
RCT 1 includes all ICU admitted patients in need for ventilation and reports the number of immune compromised patients is 10%
RCT 2 includes all ICU admitted patients in need for ventilation, but does not report the number of immune compromised patients.
The number of immune compromised patients is fairly high, however RCT 2 does not report the number of included immune compromised patients. This item scores 1 point.

Example 3. MA on the use of antibiotic prophylaxis in mechanical ventilated patients.
RCT 1 includes all ICU admitted patients in need for ventilation and reports the number of immune compromised patients is 10%.
RCT 2 includes all ICU admitted patients in need for ventilation and report the number of immune compromised patients is 6%.
In this example, the absolute difference is 4%, therefore score 0 points.

2.Domain Setting heterogeneity

This domain assesses the difference in setting of the included trials.

Signalling question 1: Is there a difference in setting between the trials, such as the years the RCTs were reported?
If the conduct of the studies differs more than 15 years, score 1 point. If the conduct of the studies is not reported, use the publication year.

Signalling question 2: Was the study conducted in a high- versus -low-middle income countries? Or in other words, is it assumable the level of 'standard care' provided to the patients is the same in the RCTs?

- Score 0: If it is assumable that standard care in the included studies is the same.
- Score 1: If it is not assumable that the standard care is the same.

Signalling question 3: Were the studies conducted in the same type of unit/facility?

- Score 0: If there are slight variations in the unit or facility type and there is low risk of affecting other fields of heterogeneity.
- Score 1: If there are slight variations in the unit or facility type, but there is risk of affecting other fields of heterogeneity.
- Score 2: If two or more signalling questions have scored points and the overall setting heterogeneity is assumedly high.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 is conducted in 1990 in Denmark.
RCT 2 is conducted in 2017 in Denmark.
The conduct of the RCTs differ more than 15 years. The standard care has changed in these years. This MA scores 2 points.

Example 2. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 is conducted in 2016 in Denmark.
RCT 2 is conducted in 2017 in Uganda.
The standard care will probably differ. This MA will score 1 point. However, if we change Uganda to a large city in India, the standard care may probably be the same.

Example 3. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 is conducted in 2016 in the mixed ICU of University Hospital with 30 ICU beds in Denmark.
RCT 2 is conducted in 2017 in a medical ICU of a small town hospital with 3 ICU beds in the Netherlands.
The unit type is the same. However, probably the patients admitted to the ICU and the standard care may differ between the two groups, therefore this item scores 1 points.

3.Domain Intervention heterogeneity

This domain is defined by multiple subdomains namely;

1. Intensity, strengths, or duration of intervention
2. Timing of the intervention
3. Control intervention
4. Co-interventions

1. Intensity, strengths, or duration of intervention

This subdomain assesses the intervention used in the different RCTs.

Signalling question 1: Do all RCTs use the same intervention?

- Score 2: If different types of interventions are used, there is clearly a heterogeneity.

Signalling question 2: If one intervention is used in all RCTs, is the intervention similar in each study?
Is the dose, strength, cut-offs or duration of intervention similar?

- Score 0: When little variations, <20%, are present.
- Score 1: When slight variations, 20-30%, are present.
- Score 2: When considerable variations, > 30%, are presents.

If signalling question 2 cannot be answered by the information in the RCTs, this item should score a 1.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 uses a dopamine dosage of 1 mcg/kg/min
RCT 2 uses a dopamine dosage of 3 mcg/kg/min
In this case the dosage does differ > 30% and this item scores with 2 points.

Example 2. MA on the use of honey for wound treatment.

RCT 1 uses honey, undefined dosage.

RCT 2 uses honey 5 g/20 cm2.

The dosage is not defined, thus this item scores 1 point.**2. Timing of the intervention**

This subdomain assesses the timing of the intervention and should be scored whether or not the intervention is started at the same time.

Signalling question 1: Are the criteria of the start of the therapy well defined?

- Score 1: If there is no information.
- Score 2: If different patient groups are included.

Signalling question 2: Is the definition of the intervention stated in the different RCTs similar?

- Score 0: If the criteria on starting the therapy are similar, or differences of timing of intervention differs ≤20%.
- Score 1: If the criteria slightly differ or the timing difference is 20-30%.
- Score 2: If other criteria are used or the timing difference exceeds 30%.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients.

RCT 1 starts the therapy when the systolic blood pressure is < 90mmHg.

RCT 2 starts the therapy when the mean arterial pressure is < 60mmHg.

In this example there is clearly a difference between the start of the therapy. Therefore, this item scores 2 points.

Example 2. MA on the start of barbiturates in traumatic brain injury patients.

RCT 1 includes patients within 2 hours after admission to the hospital.

RCT 2 includes patients within 48 hours after admission to the hospital.

There is a substantial difference in timing, therefore this item scores 2.

3. Control intervention

This subdomain assesses the use of comparable control interventions in the RCTs.

Signalling question 1: Do all RCTs use the same control intervention such as placebo, 'active' intervention, or no control?

- Score 0: If all RCTs use same control intervention or little variations, <10%, are present.
- Score 1: If all RCTs use placebo or no control or the same control intervention with slight variations, 10-20%.
- Score 2: If RCTs included in the MA use different control interventions or the same control intervention with considerable variations >20%.

The focus is on the type of control intervention and to lesser extent the dosage or timing of the control intervention., however assess as item 6 if an active control intervention is used.

Example 1. MA on the use of desmopressin for nocturia.
RCT 1 uses alfuzosin 10 mg as a control intervention
RCT 2 uses a placebo.
The control interventions are different, therefore this MA scores 2 points on this subdomain.

4. Co-interventions

This subdomain assesses the use of different co-interventions in the RCTs.

- Signalling question 1: Do the RCTs give information on standard care?

Standard care is a widely used term that should be defined by the RCTs as the standard care often differs between hospitals (even within one country).

- Score 0: If it is assumable that all RCTs will have the same standard care or if no information is given on standard care but it is assumable that the RCTs use the same standard care, this item should score a 0.
- Score 1: If no information is given on standard care and it is not assumable that the RCTs use the same standard care, this item should score a 1.

Signalling question 2: Do the RCTs state a specific co-intervention?

- Score 0: If it is assumable that the other RCTs also used this specific co-intervention as standard care or if other RCTs do not use the same co-intervention, but the effect of the co-intervention will assumedly be little.

- Score 1: If other RCTs use the same co-intervention, but with slight variation (<30 %).
- Score 2: If it is not assumable that the co-intervention is usual care, or differences in use of co-interventions ≥30 %.

Signalling question 3: Is the described 'usual care' usual care?

- Score 2: If co-interventions are used that should not be considered 'usual care'.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 states patients received the usual care.
RCT 2 states the use of fluid resuscitation and oxygen therapy.
In this example standard care of septic shock includes fluid resuscitation and oxygen therapy. Therefore, this example scores 0 points if there is no indication that standard care differed substantially on other interventions

Example 2. MA on the use of dopamine for blood pressure regulation in septic shock patients.
RCT 1 states patients received the usual care.
RCT 2 states all patients were administered hydrocortisone.
The use of hydrocortisone in septic shock patients is not standard care, therefore this item scores 2 points.

Example 3. MA on prophylactic antibiotics in ventilated patients.
RCT 1 includes post-operative liver transplantation patients.
RCT 2 includes post-operative cardiac surgery patients.
The patients included in RCT 1 will also receive immunosuppressive medication, therefore the co-interventions will differ between RCT 1 and 2. This item scores 2 points.

4.Domain Outcome heterogeneity

This domain is defined by two categories:

1. Definition of the outcome
2. Timing of outcome measurement

1. Definition of the outcome

Signalling question 1: Is the definition of the outcome in the meta-analysis and the RCTs similar?

- Score 0: If the same definition or *criteria* used in the RCTs and meta-analysis are the same.
- Score 1: If there are slight variation in the definition of the outcome.
- Score 2: If there are considerable variation in definition of outcome.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients has the outcome all-cause mortality.
In this case mortality is not disputable; a patient is alive or deceased. This MA scores 0 points on this item. If the outcome differed among the RCTs and included both disease-specific mortality, e.g. mortality in organ-confined bladder cancer, and non-organ-confined mortality the MA would score 1.

Example 2. MA on the use of antibiotic prophylaxis in mechanical ventilated patients has the outcome pneumonia.
RCT 1 defines pneumonia as positive sputum cultures.
RCT 2 defines pneumonia as diagnosed by a radiologist on a x-ray.
The definitions between the RCTs differ a lot, thus the MA scores 2 points on this item.

Example 3. MA on the use of antibiotic prophylaxis in mechanical ventilated patients has the outcome pneumonia.
RCT 1 defines pneumonia as one positive sputum culture.
RCT 2 defines pneumonia as at least two positive sputum cultures.
The definitions slightly variate between trials, The MA scores 1 point.

2. Timing of outcome measurement

Signalling question 1: Is the time of the outcome measurement the same in all RCTs?

- Score 0: If the difference in the follow up of the outcome is less than one month.

- Score 1: If the difference in follow up is more than 1 month, but less than or equal to 3 months, or if timing of outcome measurement is not reported
- Score 2: If the difference in follow up exceeds 3 months.

Example 1. MA on the use of dopamine for blood pressure regulation in septic shock patients has the outcome mortality.
RCT 1 scores mortality at day 14 after start of the therapy.
RCT 2 score mortality at 6 months after start of the therapy.
The difference between the follow up exceeds 3 months, thus this MA scores 2 points.

Example 2. MA on the use of dopamine for blood pressure regulation in septic shock patients has the outcome mortality.
RCT 1 scores mortality at day 14 after start of the therapy.
RCT 2 scores mortality at 28 days after start of the therapy
The difference is less than 1 month, therefore score 0 points.

Acknowledgement

We thank The Innovation Fund Denmark for providing a grant to Center For Research in Intensive Care (CRIC) which have made it possible for Copenhagen Trial Unit (CTU) as a partner of CRIC to write this protocol during Marija Barbateskovic PhD study.

References

1. Garattini, S., Jakobsen, J. C., Wetterslev, J., Bertele', V., Banzi, R., Rath, A., Neugebauer, E. A. M., Laville, M., Masson, Y., Hivert, V., Eikermann, M., Aydin, B., Ngwabyt, S., Martinho, C., Gerardi, C., Szmigielski, C. A., Demotes-Mainard, J. & Gluud, C. Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. European Journal of Internal Medicine. 32: 13-21; 2016

2. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ. 1994 Nov 19;309(6965):1351-5. Review.

3. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 2011.

4. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(3):167–88.

5. Koster TM, Wetterslev J, Gluud C, Keus F, van der Horst ICC. Systematic overview and critical appraisal of meta-analyses of interventions in intensive care medicine. Submitted to Acta Anaesth Scand.

6. Adhikari NKJ, Fowler RA, Bhagwanjee S, Rubenfeld GD. Critical care and the global burden of critical illness in adults. The Lancet. 2010. vol 376 (9749) p. 1339–46.

7. Reporting and interpreting missing health-related quality of life data in intensive care trials: Protocol for a systematic review. Kjaer MN, Madsen MB, Møller MH, Egerod I, Perner A. Acta Anaesthesiol Scand. 2019 aas.13326.

8. Halpern NA. Can the costs of critical care be controlled? Curr Opin Crit Care. 2009;4(6):591–6.

9. Edbrooke DL, Stevens VG, Hibbert CL, Mann AJ, Wilson AJ. A new method of accurately identifying costs of individual patients in intensive care: the initial results. Intensive Care Med. 1997 Jun;23(6):645-50.

10. Delaney A, Bagshaw SM, Ferland A, Laupland K, Manns B, Doig C. The quality of reports of critical care meta-analyses in the Cochrane Database of Systematic Reviews: an independent appraisal. Crit Care Med. 2007 Feb;35(2):589-94

11. Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, Doig CJ. A systematic evaluation of the quality of meta-analyses in the critical care literature. Crit Care. 2005 Oct 5;9(5):R575-82.

12. Cook DJ, Mulrow CD, Haynes RB.Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997 Mar 1;126(5):376-80.

13. Imberger G. Clinical guidelines and the question of uncertainty. Br J Anaesth. 2013 Nov;111(5):700–2.

14. Savovic J, Turner R, Mawdsley D, Jones HE, Beynon R, Higgins JPT, Sterne J. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: The ROBES Meta-Epidemiologic Study. Am Jr Epidemiology 2016: 1-10.

15. Rhodes KM, Turner R, Savovic J, Jones HE, Mawdsley D, Higgins JPT. Between-trial heterogeneity in meta-analyses may be partially explained by reported design characteristics. Jr Clin Epidem 2017: 45-55.

16. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008 26;336(7650): 924–6.

17. Keus F, Wetterslev J, Gluud C, van Laarhoven CJHM. Evidence at a glance: error matrix approach for overviewing available evidence. BMC Med Res Methodol. 2010 Jan;10:90.

18. Reade MC, Delaney A, Bailey MJ, Angus DC. Bench-to-bedside review: avoiding pitfalls in critical care meta-analysis--funnel plots, risk estimates, types of heterogeneity, baseline risk and the ecologic fallacy. Crit Care. 2008;12(4):220. doi: 10.1186/cc6941.

19. Gagnier JJ, Moher D, Boon H, Beyene J, Bomabardier C. Investigating clinical heterogeneity in systematic reviews: aa methodologic review of guidance in the literature. BMC Med Res Methodol 2012, 12:111.

20. Gagnier JJ, Morgenstern H, Altman DG, Berlin J, Chang S, McCulloch P, Sun X, Moher D. Med Res Methodol 2013, 13:106.

21. Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, Gluud C, Devereaux PJ, Wetterslev J.Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. PLoS One. 2012;7(7):e39471. doi: 10.1371/journal.pone.0039471.

22. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches and applications. research in social and Administrative Pharmacy. 2013;9: 330-338.

23. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med: 2002 **21**:439–458.

24. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol. 2009 Jan;9:86.

25. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol. 2008;61(1):64–75.

26. Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C. Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. BMC Med Res Methodol. 2014 Jan;14:120.

27. TSA - Trial Sequential Analysis. [Computer program on www.ctu.dk/tsa/]. Copenhagen Trial Unit, 2011;