A new tool to assess Clinical Diversity In Meta-analyses (CDIM) of interventions

Barbateskovic, Marija; Koster, Thijs M; Eck, Ruben J; Maagaard, Mathias; Afshari, Arash; Blokzijl, Fredrike; Cronhjort, Maria; Dieperink, Willem; Fabritius, Maria L; Feinberg, Josh

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2021.01.023

Publication date:
2021

Journal of Clinical Epidemiology 135 (2021) 29–41

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A new tool to assess Clinical Diversity In Meta-analyses (CDIM) of interventions

Marija Barbateskovic[a,b,*], Thijs M. Koster[c], Ruben J. Eck[d], Mathias Maagaard[a], Arash Afshari[e], Fredrike Blokzijl[f], Maria Cronhjort[g], Willem Dieperink[c], Maria L. Fabritius[h], Josh Feinberg[i], Craig French[j], Barzi Gareb[k], Anja Geisler[l], Anders Granholm[m], Bart Hiemstra[n], Ruixue Hu[o], Georgina Imberger[p], Bente T. Jensen[q], Andreas B. Jonsson[m], Oliver Karam[r], De Zhao Kong[s,t], Steven K. Korang[a], Geert Koster[c], Baoyong Lai[u], Ning Liang[v], Lars H. Lundstrøm[w], Søren Marker[b,m], Tine S. Meyhoff[b,m], Emil E. Nielsen[i], Anders K. Nørskov[a], Marie W. Munch[m], Emilie C. Risom[x], Sofie L. Rygård[m], Sanam Safi[a], Naqash Sethi[a], Fredrik Sjövall[y], Susanne V. Lauridsen[z,aa], Nico van Bakelen[k], Meint Volbeda[c], Iwan C.C. van der Horst[bb], Christian Gluud[a], Anders Perner[b,m], Morten H. Møller[b,m], Eric Keus[c], Jørn Wetterslev[a,b]

[a] *Copenhagen Trial Unit (CTU), Centre for Clinical Intervention Research, Capital Region of Denmark, Denmark*
[b] *Centre for Research in Intensive Care (CRIC), Rigshospitalet, University of Copenhagen, Copenhagen, Denmark*
[c] *Department of Critical Care, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[d] *Department of Internal Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[e] *Department of Pediatric and Obstetric Anesthesia, University of Copenhagen, Rigshospitalet, Denmark*
[f] *Department of Cardiothoracic Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[g] *Karolinska Institutet, Department of Clinical Science and Education, Södersjukhuset, Section of Anaesthesia and Intensive Care, Stockholm, Sweden*
[h] *Department of Anaesthesiology, Centre of Head and Orthopaedics, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark*
[i] *Department of Internal Medicine, Cardiology Section, Holbaek Hospital, Holbaek, Denmark*
[j] *Centre for Integrated Critical Care, The University of Melbourne, Melbourne, Australia*
[k] *Department of Oral and Maxillofacial Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[l] *Department of Anaesthesiology, Zealand University Hospital, Koege, Denmark*
[m] *Department of Intensive Care, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark*
[n] *Department of Anesthesiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[o] *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China*
[p] *Department of Anaesthesia and Pain Medicine, Western Health, Centre for Integrated Critical Care, University of Melbourne, Melbourne, Australia*
[q] *Department of Urology, Aarhus University Hospital, Denmark*
[r] *Division of Pediatric Critical Care Medicine, Children's Hospital of Richmond at VCU, Richmond, VA, USA*
[s] *Liaoning University of Traditional Chinese Medicine, Shenyang, China*
[t] *Department of Science and Technology Management, The Affiliated Hospital of Liaoning University of Traditional Chinese Medicine, Shenyang, China*
[u] *The Third Affiliated Hospital of Beijing, University of Chinese Medicine, Beijing, China*
[v] *Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China*
[w] *Department of Anaesthesiology and Intensive Care, Nordsjællands Hospital, Hillerød, Denmark*
[x] *Department of Cardiothoracic Anesthesiology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark*
[y] *Department for Intensive- and perioperative care, Skane University Hospital, Malmö, Sweden, Department for Clinical Sciences, Mitochondrial Medicine, Lund University, Lund, Sweden*
[z] *WHO-CC, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen University Hospital, Copenhagen Denmark*
[aa] *Department of Urology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark*

Abbreviations: CDIM, clinical diversity in meta-analyses; CDIMS, clinical diversity in meta-analyses score; CI, confidence interval; GRRAS, guidelines for Reporting Reliability and Agreement Studies; ICU, intensive care unit; SPSS, statistical package for social sciences.

Conflict of interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: MB and JW report grants from Innovation Fund Denmark during the conduct of the study. TSM reports grants from the Novo Nordisk Foundation and grants from Sofus Friis Foundation outside the submitted work. AP reports two grants from the Novo Nordisk Foundation related to intensive care research. The authors report no other relationships or activities that could appear to have influenced the submitted work.

* Corresponding author.
E-mail address: marija.barbateskovic@ctu.dk (M. Barbateskovic).

https://doi.org/10.1016/j.jclinepi.2021.01.023

[bb] Department of Intensive Care Medicine, Maastricht University Medical Center +, University Maastricht, Maastricht, The Netherlands

Accepted 8 January 2021; Available online 6 February 2021

Abstract

Objective: To develop and validate Clinical Diversity In Meta-analyses (CDIM), a new tool for assessing clinical diversity between trials in meta-analyses of interventions.

Study design and setting: The development of CDIM was based on consensus work informed by empirical literature and expertise. We drafted the CDIM tool, refined it, and validated CDIM for interrater scale reliability and agreement in three groups.

Results: CDIM measures clinical diversity on a scale that includes four domains with 11 items overall: setting (time of conduct/country development status/units type); population (age, sex, patient inclusion criteria/baseline disease severity, comorbidities); interventions (intervention intensity/strength/duration of intervention, timing, control intervention, cointerventions); and outcome (definition of outcome, timing of outcome assessment). The CDIM is completed in two steps: first two authors independently assess clinical diversity in the four domains. Second, after agreeing upon scores of individual items a consensus score is achieved. Interrater scale reliability and agreement ranged from moderate to almost perfect depending on the type of raters.

Conclusion: CDIM is the first tool developed for assessing clinical diversity in meta-analyses of interventions. We found CDIM to be a reliable tool for assessing clinical diversity among trials in meta-analysis.

Keywords: Meta-analysis; Systematic review; Evidence; Quality; Heterogeneity; Diversity; Tool

What is new?

- This paper describes Clinical Diversity In Meta-analyses (CDIM), the first tool scientifically developed to assess clinical diversity in meta-analyses.

Key findings

- CDIM was developed using a rigorous methodology and consists of four domains: setting diversity; population diversity; intervention diversity; and outcome diversity.
- We observed high interrater scale reliability and agreement of the CDIM tool.
- We did not observe an association between clinical diversity assessed by CDIMS and statistical heterogeneity measured with inconsistency (I^2).

What this adds to what was known?

- Assessment of clinical diversity in meta-analyses is usually not conducted transparently and systematically. Although subgroup analyses and meta-regression analyses may detect differences in treatment effect size associated with trial characteristics, the overall clinical diversity is usually not assessed. We are not aware of any tool designed to assess clinical diversity in meta-analyses. Thus, we developed CDIM—the first tool specifically developed for assessing clinical diversity in meta-analyses.

What is the implication, what should change now?

- We encourage authors of systematic reviews to use CDIM to assess clinical diversity between the included trials.

1. Introduction

A meta-analysis of high-quality randomized clinical trials is considered the best available evidence in health care management and often forms the basis of clinical practice guidelines and for protocols of randomized clinical trials [1]. Still, undetected clinical diversity, methodological and/or statistical heterogeneity may lead to inappropriate conclusions or recommendations. Several potential sources of heterogeneity exist among trials included in systematic reviews and meta-analyses. Clinical diversity can be characterized by variability in settings, participants, characteristics of interventions and comparators, use of cointerventions, and the types and timing of outcome assessments. Methodological heterogeneity, or difference in risk of bias, is characterized by variability in trial design and quality in distinct domains. Statistical heterogeneity is characterized by variability in treatment effects between or among trials [2]. The presence and magnitude of statistical heterogeneity are associated with risk of bias and may be associated with clinical sources of diversity [3,4], arise from other unknown or unrecorded trial characteristics, or from random errors ('play of chance') due to sparse data and repetitive testing [3–7]. In the context of systematic reviews, clinical diversity can be defined as differences in the clinical characteristics of trials, which may or may not lead to variations in the pooled treatment effect estimates across trials that are not explained by bias of the included trials [3,4,8].

In contrast to methodological and statistical heterogeneity [9], assessment of clinical diversity in meta-analyses is usually not conducted in a transparently and systematically [5,7]. Although subgroup analyses and meta-regression analyses may detect differences in treatment effect sizes associated with trial characteristics, the overall clinical di-

versity is usually not assessed and mapped. We are not aware of any tool designed to assess and map clinical diversity in meta-analyses.

One of the main reasons to explore clinical diversity is to inform treatment decisions, eg, by identifying specific aspects of the intervention or population that might make an intervention more or less effective. It is therefore important to improve the interpretation of systematic reviews and possibly their external validity by increasing our understanding of clinical diversity. Furthermore, as methodological heterogeneity, does not include clinical differences between trials of the included interventions, such as dosage or length of follow-up, this further call for a tool to assess, map, and screen clinical diversity between trials in a meta-analysis.

Accordingly, we aimed to develop a tool for assessing and mapping clinical diversity in meta-analyses of interventions, and to test the reliability of the tool. In a supplementary exploratory analysis, we estimated the association, if any, between a summary clinical diversity in meta-analyses score and statistical heterogeneity.

2. Methods

The development and interrater scale reliability and agreement assessments of the Clinical Diversity In Meta-analyses (CDIM) tool was conducted following our pre-published protocol and reported following the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [10,11].

2.1. Development of CDIM

We constructed CDIM during a pilot phase based on consensus work informed by empirical literature and expertise by Gagnier and colleagues [12,13] (Fig. 1a): a methodologic review of guidance of the literature on clinical diversity in systematic reviews and their consensus-based recommendations for investigating clinical diversity in systematic reviews (based on the method using a modified Delphi technique with three phases: 1. pre-meeting item generation; 2. face-to-face consensus meeting; and 3. post-meeting feedback).

One author drafted the CDIM tool which was reviewed by the author/project group and revised according to comments and circulated three times (Fig. 1a). Initially, a complete list of Cochrane reviews within the field of intensive care medicine was created [14,15]. Two authors scored the first three meta-analyses with subsequent adjustment of the CDIM tool and wrote a draft manual providing guidance on the use of CDIM. The manual was circulated between the authors and revised. Hereafter, two authors scored the next five meta-analyses from the same list and the overall summary CDIM score (CDIMS) was categorized into low, moderate, or high clinical diversity. A final version of the CDIM tool was produced to be evaluated for reliability. In the following CDIM is used for the mapping tool of Clinical Diversity In Meta-analyses while CDIMS is the summary of item scores in the CDIM tool.

2.2. Testing of CDIM

A sample of 60 meta-analyses was deemed sufficient to evaluate CDIM as 10-20 evaluations per category is considered sufficient to accurately estimate the coefficients of a regression model [16] and two times the squared amount of categories (2•categories^2) to approximate a normal distribution to be used for the analysis of quadratic weighted kappa [17].

We applied CDIM to 60 meta-analyses with a dichotomous primary outcome with at least three randomized clinical trials included (Fig. 1b). We selected in a consecutive order 20 titles (which had not already been used in the development of the CDIM) from the list of Cochrane reviews within the intensive care setting. Another 20 Cochrane reviews of interventions focusing on clinical scenarios outside the intensive care setting were selected to cover a wide range of non-intensive care interventions. These were picked by browsing The Cochrane Database of Systematic Reviews by topic. Finally, a convenience sample of 20 mainly non-Cochrane reviews with meta-analyses, of which around half were within the field of intensive care, were selected.

We evaluated CDIMS for interrater scale reliability by CDIMS scoring the 60 meta-analyses [11]. Two independent raters involved in the development of CDIMS (co-developers) and two independent raters not involved in the development of CDIMS and neither in the meta-analyses (non-developers) scored the same 40 meta-analyses. Finally, the sample of 20 mainly non-Cochrane reviews with meta-analyses was CDIMS scored by two of the review's original authors.

The two non-developers of CDIM and the 20 pairs of original review authors were instructed only by reading the guidance document – no additional guidance was given.

After individual and independent scoring of clinical diversity using CDIMS, the raters pairwise agreed upon each item score, thereby achieving summarized consensus CDIMS.

According to our protocol [10], we calculated the interrater agreement of CDIMS assessed by two independent raters in 60 meta-analyses, analyzed by linear regression and weighted Kappa for agreement between two raters, with 95% confidence intervals, of the CDIMS, with unweighted items scores, of the clinical diversity in 60 meta-analyses with clinical diversity low (score 0–10), moderate or unclear (score 11–18), and high (score 19–22):

1) We analyzed the interrater agreement of CDIMS, assessed by two independent raters involved in the development of the CDIM and the CDIM manual, in 40 meta-analyses, 20 ICU meta-analyses and 20 non-ICU

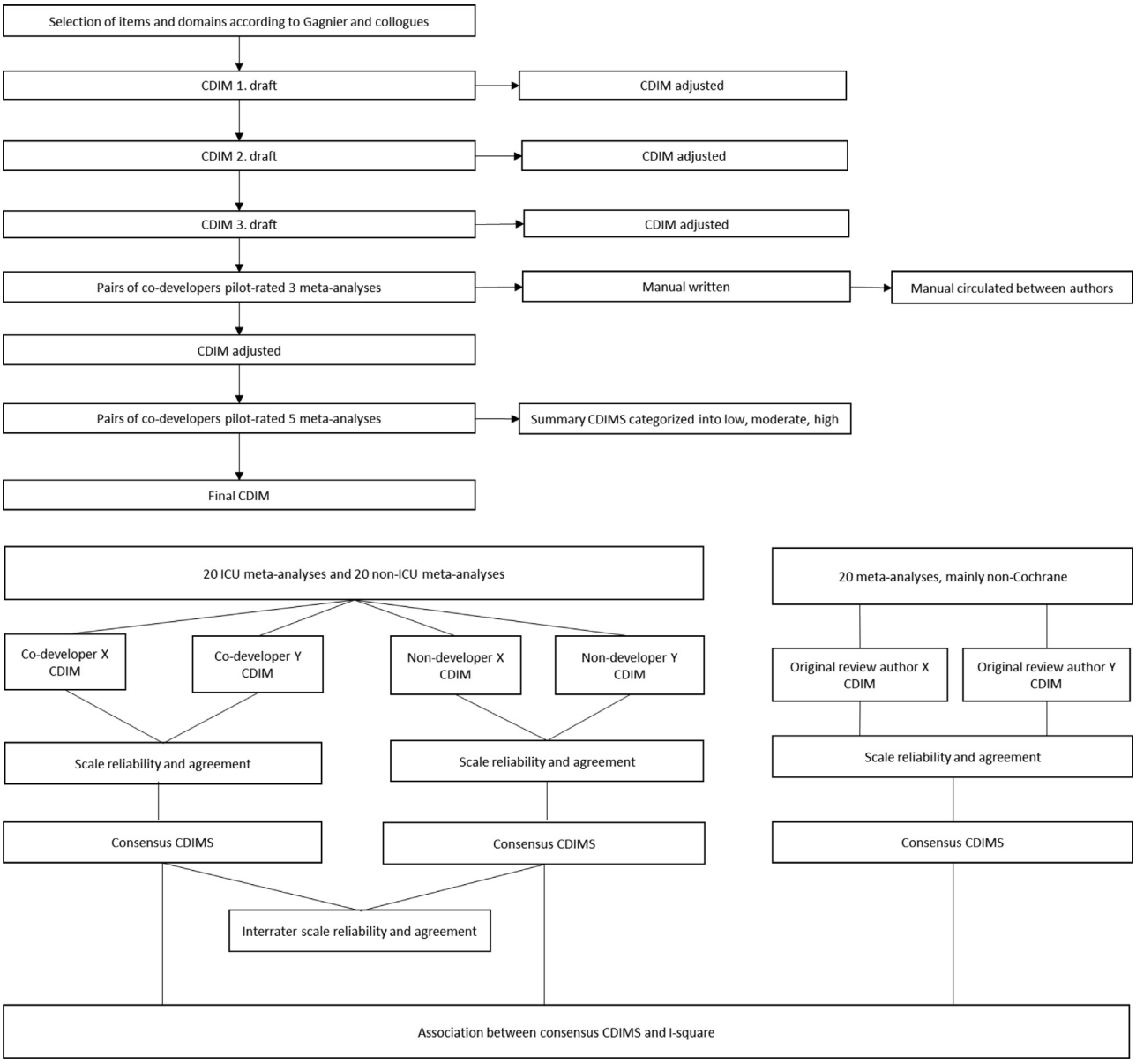

Fig. 1. (a) Process of the development of Clinical Diversity In Meta-analyses (CDIM). (b) Interrater scale reliability and agreement testing of Clinical Diversity In Meta-analyses Score (CDIMS).

meta-analyses, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression. We investigated the interaction between the interrater agreement or ICC and whether the meta-analyses were ICU or non-ICU meta-analyses.

2) We analyzed the interrater agreement of CDIMS, assessed by two independent raters not involved in the development of the CDIM and the CDIM 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. We investigate the interaction between interrater agreement or ICC and whether the meta-analyses were ICU or non-ICU meta-analyses.
3) We analyzed the interrater agreement, in 20 systematic reviews within 20 pairs of review authors scoring a meta-analysis of the primary, dichotomous outcome of their systematic review, estimating weighted Kappa and intraclass correlation coefficient (ICC) by linear regression. We investigated the interaction between the interrater agreement or ICC and whether the meta-analyses were ICU or non-ICU meta-analyses. For pairs of original review authors, we also analyzed interrater reliability within the specific domains of CDIM.

4) We analyzed the interrater agreements from step 1), 2), and 3) with 95% confidence intervals, estimating weighted Kappa and ICC by linear regression for agreement between two raters of the CDIM, with unweighted items scores, of the clinical diversity in 40 meta-analyses with low CDIMS (score 0–10), moderate or unclear CDIMS (score 11–18), and high CDIMS (score 19–22). We investigated the interrater agreement in the ICU meta-analyses and the non-ICU meta-analyses.
5) In a supplementary exploratory analysis, we estimated the agreement with weighted Kappa, between low, moderate, and high statistical heterogeneity (low $I^2 \leq 30\%$; moderate $I^2 > 30\%$ to $\leq 60\%$; high $I^2 > 60\%$) modified from Higgins et al. [18]) and low, moderate or unclear, and high clinical diversity.

We stratified the analyses of interrater scale reliability between co-developers of CDIM and non-developers of CDIM according to meta-analyses of intensive care unit (ICU) interventions or non-ICU interventions. We analyzed the possible difference between the distributions of consensus CDIMS in ICU and non-ICU meta-analyses using the Mann-Whitney test, presenting box and whiskers plots with medians, interquartile ranges, and full ranges.

The interrater reliabilities of the overall summarized CDIMS were analyzed with ICC using one-way random reliability analysis of exact agreement on average CDIMS and for single measures (single meta-analysis) for co-developers and non-developers of CDIM. A two-way random reliability analysis of exact agreement was used for pairs of original review authors. For pairs of original review authors, we also analyzed interrater reliability within the domains of CDIM.

Quadratic weighted kappa values for the agreement between the protocolized categorical classification of CDIMS (low: 0–11; moderate 12–18; high 19–22), defined after a pilot scoring, were calculated. Moreover, quadratic weighted kappa values for the agreement between the protocolized categorical classification of CDIMS and the categorical classification of I^2 in the meta-analyses (low $I^2 \leq 30\%$; moderate $I^2 > 30\%$ to $\leq 60\%$; high $I^2 > 60\%$) modified from Higgins et al. were calculated [18]. Imputed relative distances between ordinal categories in the calculation of the quadratic weighted kappa were set to one.

Additionally, linear regression analyses were performed for any associations between the raters' summarized total CDIMS. Finally, we analyzed the possible association between the consensus CDIMS and I^2 in 60 meta-analyses using linear regression. Pearson's correlation coefficients, R^2, and P-values for the linear regression coefficients being equal to zero were calculated. We plotted regression lines and regression standardized residuals including P-P plots to investigate whether residuals were normally distributed as required for a linear regression models to be adequate.

Agreement was classified as suggested by Landis and Koch: values less than 0 indicated poor, 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement [19]. All ICC and kappa values are presented with 95% confidence interval (CI). SPSS version 17 (SPSS Statistics for Windows, Chicago: SPSS Inc.) was used for the analysis of scale reliability. http://vassarstats.net/kappa.html was used to calculate kappa values.

3. Results

CDIM measures clinical diversity on an ordinal scale that includes four domains with 11 items overall, covering essential domains describing clinical diversity [12,13] (Table 1).

- The first domain aims to detect *setting diversity* by assessing differences between trials in: time of conduct; type of country development status; localization within the health care system.
- The second domain aims to detect *population diversity* by assessing differences between trials in: age; sex; patient inclusion criteria and baseline disease severity; comorbidities.
- The third domain aims to detect *intervention diversity* by assessing differences between trials in: intervention intensity/strength/duration (dose, frequency, duration, device, cut-off values); timing of intervention(s); diversity of control-interventions; use of cointerventions.
- The fourth domain aims to detect *outcome diversity* between trials by assessing differences between trial in: outcomes definitions and timing of outcome assessment(s).

The CDIM tool is used in two steps:

1. Map and screen clinical diversity in each of the four domains (11 items) (Table 1). The 11 items are each scored as low diversity (0 points), moderate (or unknown/undescribed) diversity (1 point), or high diversity (2 points), with a total range of 0–22 with the equal weight assigned to each item. The thresholds should be used as guidance only, and assessors may use other thresholds if that better suits the clinical field that is investigated. When assessors are in doubt, then assessors have the possibility of choosing the score 1 which corresponds to "unknown/undescribed/not applicable". Guidance on how to score each item is provided in the CDIM manual (Supplementary Appendix A).
2. Sum the item scores into summary CDIMS.

3.2. Interrater scale reliability and agreement of CDIMS

Four raters independently applied CDIM to 20 meta-analyses of ICU-interventions and 20 meta-analyses of non-ICU interventions, for a total of 160 evaluations. Twenty pairs (of 35 different raters) of original review authors applied CDIM to 20 meta-analyses, for a total of 40 evaluations (Supplementary Appendix B). In total, 721 tri-

Table 1. The Clinical Diversity In Meta-analyses (CDIM) tool

Domains of diversity	Items	Score	Explanation of score for extreme differences between trials in a meta-analysis
Setting	1. Years reported (A), performed in developed vs developing country (B), unit type (C)	0	No differences: A) years reported differ $<$ 15, B) No developed vs developing countries, AND C) slight variations in the unit or facility type and there is low risk of affecting other fields of diversity
		1	Slight variation (at least one of A-C involved): A) years reported differ $\geq$ 15, OR B) developed vs developing countries, 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 diversity
		2	Considerable variation (all of A-C involved): A) years reported differ $\geq$ 15, AND B) developed vs developing countries, AND C) treating units not similar (all of A-C involved), OR if the trials in the opinion of the assessor differs markedly in setting diversity
Population	2. Age	0	Mean/median age $\leq$ 10 years difference
		1	11–20 years difference in mean/median age
		2	Mean/median age $>$ 20 years difference
	3. Sex	0	% women $\leq$ 20% absolute difference between trials
		1	21%–30% absolute difference of % women between trials
		2	More than $>$ 30% absolute difference between trials
	4. Participant inclusion criteria and baseline disease severity	0	Different trials include patients that are equally ill or the difference in risk OR score for disease severity of patients $\leq$ 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)
	5. Comorbidities	0	Difference in frequency of important comorbidities $\leq$ 20% OR no comorbidities are reported in the included trials AND differences in comorbidities are assumed absent
		1	Slight differences in important comorbidities, between 21% and 30%, OR no comorbidities are reported in the trials, but differences in comorbidities are assumed
		2	Differences in frequency of important comorbidities $>$ 30% OR highly likely variations in comorbidities
			Use absolute difference when comparing important comorbidities
Intervention	6. Intensity, strengths, or duration of intervention	0	Little variation: differences in dose, strengths, devices, cut-offs, OR duration of interventions $\leq$20%
		1	Slight variation: 21% to 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 included trials
		2	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	Criteria for starting the intervention are similar, OR relative differences of timing of intervention differ $\leq$ 20%
		1	Criteria for starting the intervention differ slightly, OR the relative timing difference is 21% to 30%
		2	Criteria for starting the intervention differ, OR relative timing difference exceeds $>$ 30%

(*continued on next page*)

Table 1 (*continued*)

Domains of diversity	Items	Score	Explanation of score for extreme differences between trials in a meta-analysis
	8. Control intervention	0	All control interventions are the same
		1	Control interventions include placebo and no intervention, assess as item 6 if an active intervention is used
		2	Including trials with different active control interventions OR trials with active and placebo/no intervention
	9. Cointerventions	0	No apparent differences in cointerventions OR standard care is not described or assumed to be the same, OR equally applied in groups, OR different cointerventions are used, but the effects of the cointerventions are assumed to be small
		1	Slight variation in cointerventions OR the same cointerventions are used with slight variation (< 30% difference in, eg, doses or numbers of participants using the cointervention)
		2	Considerable differences if it is assumed that the cointervention is not usual care, OR differences in use OR, eg, doses of cointerventions > 30%
			Use absolute difference when assessing cointerventions
Outcome	10. Definition of the outcome in the meta-analysis	0	Same definition of outcome
		1	Slight variations in definition of outcome
		2	Considerable variations in definition of outcome
	11. Timing of outcome measurement	0	Less than one month between follow-up of outcome
		1	More than one but less than or equal to 3 months between follow-ups
		2	More than 3 months between follow-up of outcome

als were included in the 60 meta-analyses assessed with a median of 8 (interquartile range 5–15) trials per meta-analysis.

Main characteristics of the meta-analyses evaluated, their reference, and Supplemental Figures are presented in the Supplementary Appendix C.

CDIMS varied between 0 and 21 points in the 60 meta-analyses. Average CDIMS for all raters varied between (mean ± SD) 11.5 ± 5.4 and 14.2 ± 3.9 and the difference between average CDIMS for pairs of raters ranged between 0.3 and 2.4 (Table 2).

3.2.1. Co-developers of CDIM

Interrater scale reliability of CDIMS was almost perfect for two co-developers of CDIM with an ICC of 0.85 (95% confidence interval 0.72–0.92) for average measures and substantial with an ICC of 0.74 (0.56–0.85) for single measures. Pearson's correlation coefficient was 0.76 (0.53–0.98). Quadratic weighted kappa values for the agreement between categorical CDIMS for two co-developers was substantial with a kappa of 0.61 (0.18–1.00). Consensus CDIMS score between developers of CDIM stratified for ICU and non-ICU meta-analyses were median 18 (range 9–20) and median 12 (range 7–18), respectively ($P = 0.001$, Mann-Whitney test for different distributions of CDIMS; Supplementary Appendix C). The interrater scale reliability between two developers of CDIM in ICU meta-analyses and non-ICU meta-analyses were almost perfect as well (Table 3).

3.2.2. Non-developers of CDIM

Interrater scale reliability for two non-developers of CDIM was substantial with an ICC of 0.74 (0.51–0.86) for average measures and moderate for single measures with an ICC of 0.59 (0.34–0.76). Pearson's correlation coefficient was 0.72 (0.56–0.88). Quadratic weighted kappa values for the agreement between categorical CDIMS for two non-developers was moderate with a kappa of 0.41 (0.14–0.69). Consensus CDIMS between non-developers of CDIM stratified for ICU and non-ICU meta-analyses were median 17 (range 7–21) and median 12 (range 5–19), respectively ($P = 0.016$, Mann-Whitney test for different distributions of CDIMS; Supplementary Appendix C). The interrater scale reliability between two non-developers of CDIM on average measures in ICU meta-analyses and non-ICU meta-analyses were substantial and moderate, respectively (Table 3), and moderate and fair, respectively for single measures (Table 3).

3.2.3. Pairs of original review authors

Interrater scale reliability of CDIMS for two original review authors was almost perfect with an ICC of 0.94 (0.85–0.98) for average measures and 0.89 (0.75–0.96) for single measures. Pearson's correlation coefficient was 0.90

Table 2. Interrater agreements of Clinical Diversity In Meta-analyses Score (CDIMS) stratified for types of raters as developers, original review authors, and non-developers of CDIM

Coefficients Datasets analyzed	Scale reliability: intraclass correlation coefficient on average measures (95% CI)	Intraclass correlation coefficient on single measures (95% CI)	Pearson's correlation coefficient (95% CI)	R^2, *P*-value for test of linear regression coefficient equal to 0, and model fit	Constant (95% CI) in linear regression equation, raters mean ± SD	Quadratic weighted kappa (95% CI) for agreement between low, moderate, or high CDIMS[a]
Two co-developers of CDIM[b]	0.85 (0.72–0.92)	0.74 (0.56–0.85)	0.76 (0.53–0.98)	0.54 $P < 0.0001$ Residual plots suggest goodness of model fit	3.0 (-0.21–6.2) 1. Rater: 13.6 ± 3.6 2. Rater: 13.3 ± 3.7	0.61 (0.18–1.00)
Pairs of original review authors[c]	0.94 (0.85–0.98)	0.89 (0.75–0.96)	0.90 (0.69–1.12)	0.82 $P <$ 0.0001 Residual plots suggest goodness of fit	0.13 (-2.9 to 3.2) 1. Rater: 13.2 ± 6.2 2. Rater: 12.1 ± 6.2	0.72 (0.42–1.00)
Two non-developers of CDIM[b]	0.74 (0.51–0.86)	0.59 (0.34–0.76)	0.72 (0.56–0.88)	0.52 $P <$ 0.0001 Residual plots suggest goodness of model fit	7.9 (5.8–10.0) 1. Rater: 13.9 ± 3.9 2. Rater: 11.5 ± 5.4	0.41 (0.14–0.69)
Consensus scores from co-developers and non-developers of CDIM[b]	0.91 (0.83–0.95)	0.84 (0.72–0.91)	0.85 (0.81–1.22)	0.73 $P <$ 0.0001 Residual plots suggest goodness of model fit	0.75 (-3.7 to 2.2) 1. Rater: 14.2 ± 3.9 2. Rater: 13.7 ± **4**.5	0.68 (0.38–0.98)

[a] Low CDIMS: 0 to 10; Moderate CDIMS: 11 to 18; high CDIMS: 19 to 22.
[b] One-way random reliability analysis of exact agreement analysis of 40 meta-analyses rated with CDIMS.
[c] Two-way random reliability of exact agreement analysis of 20 pairs of raters of 20 meta-analyses not involved in the development of CDIMS.
CI is confidence interval.

Table 3. Interrater agreements of Clinical Diversity in Meta-analyses Score (CDIMS) stratified for ICU and non-ICU meta-analyses

Coefficients Datasets analyzed	Scale reliability: intraclass correlation coefficients on average measures (95% CI)	Intraclass correlation coefficients on single measures (95% CI)	Pearson's correlation coefficient (95% CI)	Raters' means ± SD
ICU meta-analyses Interrater agreement[a] between two co-developers of CDIM[a]	0.71 (0.29–0.89)	0.55 (0.17–0.80)	0.54 (0.12–0.89)	1. Rater: 15.3 ± 3.2 2. Rater: 15.7 ± 3.0
Non-ICU meta-analyses Interrater agreement[a] between two co-developers of CDIM	0.82 (0.56–0.93)	0.70 (0.39–0.87)	0.75 (0.35–0.91)	1. Rater: 12.0 ± 3.2 2. Rater: 11.0 ± 2.7
ICU meta-analyses Interrater agreement[a] between two non-developers of CDIM	0.78 (0.45–0.91)	0.64 (0.29–0.84)	0.69 (0.27–0.86)	1. Rater: 15.6 ± 3.9 2. Rater: 14.1 ± 4.7
Non-ICU meta-analyses Interrater agreement[a] between two non-developers of CDIM	0.55 (-0.13 to 0.82)	0.38 (-0.06 to 0.69)	0.63 (0.17–0.70)	1. Rater: 12.4 ± 3.4 2. Rater: 9.0 ± 4.9

[a] One-way random reliability analysis of exact agreement in 20 meta-analyses rated with CDIMS. CI is confidence interval. SPSS version 17 was used.

(0.69–1.12) (Fig. 2). Quadratic weighted kappa values for the agreement between two original review authors was substantial with a kappa of 0.72 (0.42–1.00).

Interrater scale reliability of CDIMS for two original review authors on the four CDIM domains was consistent with a scale reliability ranging from substantial to almost perfect across domains. The domain summary scale reliability ranged from 0.68–0.93 on average measures and from 0.51–0.87 for single meta-analyses (Supplementary Appendix 3).

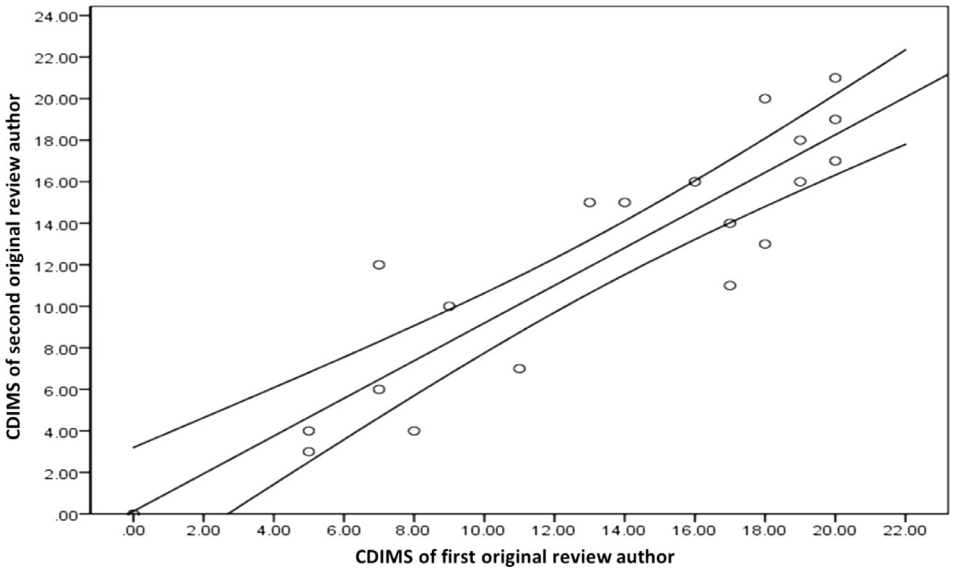

Fig. 2. Fitted regression line (Y = 0.90 • X + 0.13) of Clinical Diversity In Meta-analyses Score (CDIMS) from second original review author on CDIMS from first original review author in 20 meta-analyses from mainly non-Cochrane reviews. Hyperbolic lines around fitted line represent 95% CI for the regression line. R^2= 0.82.

3.2.4. Consensus scores between developers and non-developers of CDIM

Interrater scale reliability of consensus CDIMS between developers and non-developers of CDIMS was almost perfect with an ICC of 0.91 (0.83–0.95) for average measures and 0.84 (0.72–0.91) for single measures. Pearson's correlation coefficient was 0.85 (0.81–1.22) (Supplemental Appendix C). Quadratic weighted kappa values for the agreement between the categorical consensus CDIMS was substantial with a kappa of 0.68 (0.38–0.98).

Linear regression showed that a linear model explained from 52%–82% of the covariation in CDIMS between raters regardless of the meta-analyses being ICU or non-ICU meta-analyses (Table 2). Model of fit analyses justified a linear regression model as standardized residuals were normally distributed.

3.2.5. Association between clinical diversity expressed as consensus CDIMS and statistical heterogeneity expressed as I^2

Consensus CDIMS from both developers and non-developers of CDIMS supplemented with consensus CDIMS for pairs of original review authors indicated an absence of association with regression coefficients close to zero with narrow CIs: -0.02 (-1.6–1.4) and -0.13 (-2.0 to 0.7), respectively (Table 4 and Supplementary Appendix C). In fact, a linear model seems unjustified, as analyses of standardized residuals indicated the absence of a normal distribution. Quadratic weighted kappa values for the agreement between categorical consensus CDIMS and categorical statistical heterogeneity was not calculable because the observed concordance was smaller than mean chance concordance (Table 4).

4. Discussion

We aimed to create a systematic approach to assess and map differences in clinical characteristics (diversity) of included trials in a meta-analysis. By clinical characteristic differences, we mean factual clinical differences between or among the included trials in a meta-analysis. To assist in this aim, we developed CDIM and its summary score CDIMS, a mapping and screening tool developed to characterize the factual clinical diversities present in the meta-analysis which may or may not have been explored for effect modifying properties in a meta-analysis from a systematic review.

We evaluated CDIM in three groups of assessors. The highest interrater scale reliability and agreement on both average and single summarized measures of CDIMS and categorical classification of CDIMS (low, moderate, high) were achieved in groups of original review authors. Co-developers achieved lower interrater scale reliability and agreement compared to original review authors. Non-developers of CDIM who were not involved in the rated meta-analyses achieved the lowest interrater reliability and agreement. Although interrater scale reliability and agreement between non-developers of CDIM were only moderate to substantial for average measures, single measures, and categorical classifications of CDIMS, respectively, the reliability and agreement increased to substantial and almost perfect, respectively, when either scores from two co-developers of CDIM or two original review authors were compared. Even for individual domains, the reliability and agreement were substantial to almost perfect when ratings of two original review authors were compared. The external reliability (or generalizability) tested by assessing consensus scores from the group of co-developers and

Table 4. Regression of consensus Clinical Diversity In Meta-analyses Score (CDIMS) on statistical heterogeneity (I^2) and Kappa between categorized CDIMS and categorized I^2.

Coefficients Dataset analyzed	Regression coefficient (95% CI)	R^2, P-value for test of regression coefficient equal to 0, and model fit	Constant (95% CI) in linear regression equation	Quadratic weighted kappa (95% CI) for agreement between low, moderate, or high CDIMS[a] and low, moderate and high statistical heterogeneity
Consensus CDIMS (from two co-developers)[b] versus I^2[c]	-0.02 (-1.6 to 1.4)	0.000 $P = 0.88$ Residual plots suggest lack of model fit	21.5 (-0.07 to 43.1)	Kappa is not calculated for this data set because observed concordance is smaller than mean-chance concordance
Consensus CDIMS (from two non-developers)[b] versus I^2[c]	-0.13 (-2.0 to 0.7)	0.016$P =$ 0.34 Residual plots suggest lack of model fit	28.7 (9.2–48.2)	Kappa is not calculated for this data set because observed concordance is smaller than mean-chance concordance

Regression analysis of consensus CDIMS and I^2 in 60 meta-analyses rated with CDIM.
[a] Low CDIMS 0–10; moderate CDIMS 11–18; high CDIMS 19–22.
[b] Supplemented with consensus scores from original review authors.
[c] Low $I^2 \leq 30\%$; moderate $I^2 > 30\%$ to $\leq 60\%$; high $I^2 > 60\%$. NC is not calculated. SPSS version 17 was used.

non-developers of CDIM was almost perfect when analyzing interrater scale reliability and substantial when analyzing CDIMS categories stressing the fact that consensus is important to achieve when assessing clinical diversity with CDIMS. Consensus scores of co-developers and non-developers showed significant higher CDIMS within intensive care meta-analyses compared to non-ICU meta-analyses.

Moreover, we observed absence of a linear association between clinical diversity measured as CDIMS and statistical heterogeneity quantified by I^2 as regression coefficients were close to zero with narrow confidence intervals. To summarize our exploratory analyses, it appears that clinical and statistical heterogeneity are two different aspects of heterogeneity in meta-analyses. The results showed that we cannot expect statistical heterogeneity to be high just because clinical diversity is, or vice versa. In fact, several meta-analyses with zero I-square had high CDIMS. In theory this makes sense and may be explained by: 1) The amount of data in the meta-analysis may be too small to reveal statistical heterogeneity despite abundant clinical diversity (CDIMS). On the other hand, given high precision on the effect estimates in the included trials (eg, with continuous effect measures), a high I-square/D-square may not represent genuine effect heterogeneity [20]; 2) The difference in clinical diversity between trials does not necessarily result in different treatment effects (high statistical heterogeneity) due to an overall similar treatment effect regardless of eg, age of included populations and dose of the examined intervention; 3) In situations where trials in a meta-analysis are similar across CDIM items that are hypothesized to be the most important effect modifiers, but different across other factors, eg, risk of bias, which are not covered by CDIM; 4) Other factors such as reporting bias and other bias may influence I-square and not CDIMS.

4.1. Strengths and limitations

Our approach used in the development of CDIM has several strengths. We relied strongly on the consensus reports and expert panel from which the items and domains covered in CDIM originate [12,13]. The CDIM tool was developed over several steps and a final version of the CDIM tool was extensively evaluated in a relatively large sample of meta-analyses of different settings, populations, interventions and outcomes by three groups of raters. It includes a domain and item-based approach supported by signaling questions in a manual similar to other tools used in the systematic review process [2,21,22] and it appears that the raters found the CDIM tool operational determined by the fact that the developers only received two clarifying questions among non-developers and original review authors.

Knowledge about the medical field and interventions assessed in the meta-analyses seems preferable when assessing clinical diversity with CDIM; other expertise such as knowledge of trial methodology or statistics is not required. Application of the tool requires some time investment as full trial reports from all trials included in a meta-analysis must be explored carefully, especially when many trials with low clinical diversity in one or more domains are included in the assessed meta-analysis. Conversely, the scoring using CDIM can be completed rather quickly in the presence of high clinical diversity for an item when just two trials differ substantially (see manual, Supplementary Appendix A). Nonetheless, we recommend looking for the specific information needed to assess all items in all trials to get a full overview of the clinical diversity in the meta-analysis.

In some circumstances some items may partly overlap. This is the case when a meta-analysis is conducted in a

'lumping' review that includes all participants regardless of eg, age, and thus may lead to high clinical diversity between the included trials for the items of age, but also for items such as participant inclusion criteria, baseline disease severity, and comorbidities, consequently leading to possible double counts.

In our sample, clinical diversity in meta-analyses of interventions in the field of intensive care appears to be high as compared to the group of meta-analyses in other medical fields. This difference indicates higher clinical diversity in meta-analyses in the field of intensive care, but it may also be a chance finding. Nevertheless, the domains and items included in the CDIM tool have been selected to be key categories/topics especially with the purpose of investigating clinical diversity in meta-analyses regardless of the medical field [13].

A reason for the imperfect agreement between the categories low, moderate and high CDIMS may be attributable to the somewhat arbitrary cut off between these categories, which may be reflected in the analyses of the quadratic weighted kappa values.

4.2. Implications

The CDIM tool is designed to be applicable in all medical fields and intended to be used by multiple users such as researchers and guideline panels conducting or critically appraising meta-analyses.

The CDIM tool is to be used as a mapping and screening tool that should help authors of systematic reviews in a structured and transparent way to compare the PICO characteristics across trials and to the review PICO and other clinical characteristics of the included trials in the meta-analysis. Thus, the tool is only intended to be used as a mapping and screening tool that may be used to point out clinical characteristics that ought to be explored further. The scorings cannot tell how and why the meta-analytic result is affected by clinical diversity. The scoring will enable authors *post hoc* and in future updates of their review to explore factors that might modify intervention effects importantly, and in that way identify potential effect modifiers. Further, systematic reviewers may easily incorporate a plan in their protocol for a new systematic review to use CDIM to check whether possible clinical diversity has been explored sufficiently. CDIM can then be used to define and select subgroup analyses as we suggest items with a score of 2 possibly ought to be explored in subgroup analyses; this way CDIM can be used prospectively to select subgroup analyses (Fig. 3).

The summary clinical diversity measure, CDIMS, is intended to help the systematic reviewers become informed about the degree of clinical diversity between the trials; the higher the overall score, the more should this be explored guided by the mapping of the clinical differences by eg, subgroup analyses and highlighted in the systematic review or in future updates of the review. Furthermore, if CDIMS is zero or low and the items creating this have

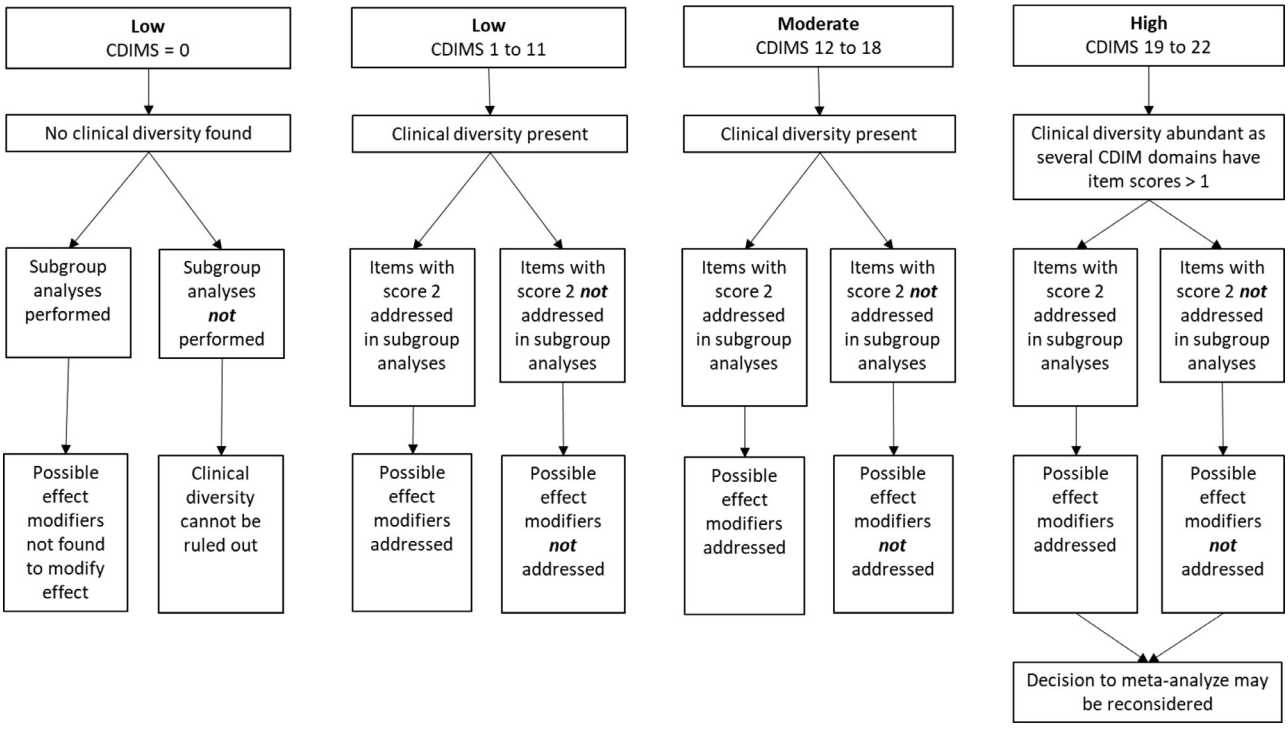

Fig. 3. Interpretation of Clinical Diversity In Meta-analyses Score (CDIMS) and how to conclude on the assessment of clinical diversity between included trials in a meta-analysis.

been explored in subgroup analyses, then clinical diversity may have been addressed or may not be a problem and effect modifiers have or have not been detected. On the other hand, if CDIMS is moderate or high and items are contributing to a high CDIMS which have not been explored in the systematic review, by subgroup analysis or meta-regression, then yet unknown effect modifiers (confounders) may still be a possibility and should be explored in the meta-analysis and in future updates of the meta-analysis. Even though a relevant PICO has been phrased in the protocol of a systematic review, clinical diversity within the panel of trials included may vary a great deal on several (other) clinical characteristics not covered by the inclusion or exclusion criteria.

The panel of items chosen in CDIM is based on the consensus panel of possible clinically important characteristics that ought to be explored in all meta-analysis and CDIM is not a panel of universally known effect modifiers. However, this does not rule out the possibility of combining knowledge from eg, large observational studies pointing to relevant risk factors or effect modifiers. If such knowledge exists and CDIM reveals clinical diversity, due to a diversity of suspected risk factors, among the included trials there is even more reason for exploring whether effect modification by these factors is present. If risk factors, derived from large observational studies has been identified, and not detected by CDIM, then we suggest that these factors should be assessed in subgroup analyses as well.

Our analyses illustrate that CDIM is a reliable mapping and screening tool for assessing clinical diversity in meta-analyses. We consider to use CDIM in the systematic review process to quantify overall clinical diversity, to highlight clinical diversity within specific domains and it may be practical when assessing indirectness and inconsistency in GRADE [5]. Other implications include the possibility of comparing CDIMS across meta-analyses and with statistical heterogeneity such as I^2 or D^2 [23]. However, our finding of lack of association between clinical diversity and statistical heterogeneity should be considered hypothesis-generating due to the limited number of investigated meta-analyses and scenarios. In any case, we recommend these results to be explored further. We encourage investigators to provide feedback and report experiences to the corresponding author.

In conclusion, CDIM is the first tool developed to assess clinical diversity in meta-analyses. Interrater scale reliability for overall CDIMSin various scenarios varied from moderate to almost perfect. Reliability was almost perfect between original review authors and between consensus scores of non-developers and co-developers of the CDIM tool. We consider CDIM a reliable tool and recommend using CDIM for the assessment and mapping of the overall clinical diversity in meta-analyses.

Author contributions

MB and JW conceived the project. MB organized, collected data, and oversaw the project. JW drafted the first version of the CDIM tool, and MB, TMK, FK, CG, MHM, ICCH and AP contributed to the development. MB, TMK, and JW drafted the guidance manual. MB and TMK pilot-tested CDIM and comprised the pair of co-developers; RE and MM comprised the pair of non-developers. 41/45 coauthors each CDIM scored at least one meta-analysis. JW performed the statistical analyses. MB and JW wrote the first draft of the manuscript. All authors reviewed, commented on the draft, and finally approved the manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. MB's and JW's contribution to this work, was supported by Innovation Fund Denmark [grant number 4108-00011B].

Supplementary materials

Supplementary Appendix A: CDIM tool and guidance document

Supplementary Appendix B: Data from the assessment of reliability and agreement of CDIMS

Supplementary Appendix C: Supplementary material for reliability and agreement analyses

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.01.023.

References

[1] Garattini S, Jakobsen JC, Wetterslev J, Bertele V, Banzi R, Rath A, et al. Evidence-based clinical practice: overview of threats to the validity of evidence and how to minimise them. Eur J Intern Med 2016;32:13–21.

[2] Higgins JPT, Green S (editors). Cochrane handbook for systematic reviews of interventions version 5.1.0 The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

[3] Savovic J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. Am J Epidemiol 2018;187(5):1113–22.

[4] Rhodes KM, Turner RM, Savovic J, Jones HE, Mawdsley D, Higgins JPT. Between-trial heterogeneity in meta-analyses may be partially explained by reported design characteristics. J Clin Epidemiol 2018;95:45–54.

[5] 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;336:924–6.

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

[7] 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;12:220.

[8] Imberger G. Clinical guidelines and the question of uncertainty. Br J Anaesth 2013;111:700–2.
[9] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;309:1351–5.
[10] Barbateskovic M, Koster TM, Keus F, Gluud C, Møller MH, van der Horst ICC, et al. 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. Copenhagen Trial Unit 2019. http://ctu.dk/media/13724/2019-protocol-chims-protocol-manual-ver-11.0_11-03-2019.pdf.
[11] Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96–106.
[12] Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol 2012;12:111.
[13] Gagnier JJ, Morgenstern H, Altman DG, Berlin J, Chang S, McCulloch P, et al. Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. BMC Med Res Methodol 2013;13:106.
[14] 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. Acta Anaesthesiol Scand 2018;62:1041–9.
[15] Koster TM, Wetterslev J, Gluud C, Jakobsen JC, Kaufmann T, Eck RJ, et al. Apparently conclusive meta-analyses on interventions in critical care may be inconclusive-a meta-epidemiological study. J Clin Epidemiol 2019;114:1–10.
[16] Gerke O, Moller S, Debrabant B, Halekoh U. Experience applying the guidelines for reporting reliability and agreement studies (GRRAS) indicated five questions should be addressed in the planning phase from a statistical point of view. Diagnostics 2018;8:E69.
[17] Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm 2013;9:330–8.
[18] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[19] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[20] Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008;8:79.
[21] Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225–34.
[22] Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
[23] 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;9:86.