Quantitative assessment of inconsistency in meta-analysis using decision thresholds with two new indices

Bernardo Sousa-Pinto[1,2], Ignacio Neumann[3], Rafael José Vieira[1,2], Antonio Bognanni[4,5,6], Manuel Marques-Cruz[1,2], Sara Gil-Mata[1,2], Simone Mordue[7], Clareece Nevill[8], Gianluca Baio[9], Paul Whaley[10,11], Guido Schwarzer[12], James Steele[13], Gavin Stewart[14], Holger J Schünemann[6,15*], Luís Filipe Azevedo[1,2]

1 – MEDCIDS - Department of Community Medicine, Information and Health Decision Sciences; Faculty of Medicine, University of Porto, Porto, Portugal

2 – CINTESIS@RISE - Health Research Network, MEDCIDS, Faculty of Medicine, University of Porto, Porto, Portugal

3 – School of Medicine, Universidad San Sebastián, Santiago, Chile

4 – Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada

5 – Department of Medicine, Evidence in Allergy Group, McMaster University, Hamilton, Canada

6 – Clinical and Epidemiology and Research Center (CERC), IRCCS Humanitas Research Hospital, & Department of Biomedical Sciences, Humanitas University, Milan, Italy

7 – Centre for Conservation Science, Cambridge, UK

8 – Department of Population Health Sciences, University of Leicester, Leicester, UK

9 – Department of Statistical Science, University College London, London, UK

10 – Lancaster Environment Centre, Lancaster University, Lancaster, UK

11 – Evidence-Based Toxicology Collaboration (EBTC), Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

12 – Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

13 – Department of Sport and Health, Solent University, Southampton, UK

14 – Evidence Synthesis Lab, School of Natural and Environmental Science, University of Newcastle, Newcastle-upon-Tyne, UK

15 – Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Allergology and Immunology, Berlin, Germany

Correspondence to: *Holger J Schünemann; Clinical and Epidemiology and Research Center (CERC), Department of Biomedical Sciences, Humanitas University & IRCCS Humanitas Research Hospital Milan, Italy; E-mail address: schuneh@mcmaster.ca

Abstract

Objective: In evidence synthesis, inconsistency is typically assessed visually and with the I^2 and the Q statistics. However, these measures have important limitations (i) if there are few primary studies of small sample sizes, or (ii) if there are multiple studies with precise estimates. In addition, with the increasing use of decision thresholds (DT), for example in GRADE Evidence to Decision frameworks, inconsistency judgments can be anchored around DTs. In this article, we developed quantitative measures to assess inconsistency based on DTs.

Study Design and Setting: We developed two measures to quantify inconsistency based on DTs – the Decision Inconsistency (*DI*) and the Across-Studies Inconsistency (*ASI*) indices. The *DI* and the *ASI* are based on the distribution of the posterior samples studies' effect sizes across interpretation categories defined by DTs. We developed these indices for the Bayesian context, followed by a frequentist extension.

Results: The *DI* informs on the *overall inconsistency of effect sizes* across interpretation categories, while the *ASI* quantifies how *different* studies are compared to each other (in relation to interpretation categories) based on absolute effects. A *DI*≥50% and an ASI≥25% are suggestive of important unexplained inconsistency. We provide an R package (metainc) and a web tool (https://metainc.med.up.pt/) to support the computation of the *DI* and *ASI*, including in the context of sensitivity analyses assessing the impact of potential uncertainty in inconsistency.

Conclusion: The *DI* and the *ASI* can contribute to quantitatively assess inconsistency, particularly as DTs are gaining recognition in evidence synthesis and health decision-making.

Key words: GRADE; Heterogeneity; Inconsistency; Meta-analysis; Systematic review

What is new

Key findings

- This study proposes two new quantitative measures to assess inconsistency in the evidence synthesis context – the Decision Inconsistency (*DI*) and the Across-Studies Inconsistency (*ASI*) indices. These indices differ from previously existing measures by considering effect size in the context of decision thresholds (DTs).
- We have developed a R package (metainc) and a web tool (https://metainc.med.up.pt/) to easily support the computation of the *DI* and the *ASI*.

What this adds to what was known

- The GRADE working group posits that inconsistency judgments can be made considering DTs. Our methods allow such judgments to be supported by quantitative indices' results.

What is the implication, what should change now?

- Quantitative assessment of inconsistency based on DTs is now possible. Therefore, judging inconsistency when assessing the certainty of evidence may now consider the quantification of the DT-related *DI* and *ASI*, alongside other approaches for appraising inconsistency.

Highlights

- GRADE assessments of inconsistency are facilitated by decision thresholds (DT)
- Two inconsistency indices have been developed to measure inconsistency based on DTs
- The new indices allow to assess the impact of uncertainty in the evidence on inconsistency

1. Introduction

In the context of evidence synthesis and appraisal, there are several methods to assess inconsistency [1]. One possibility is the visual inspection of the forest plot, which provides a simple but subjective approach. The Cochran's Q test allows for the calculation of a *p*-value, based on which it is possible to reject (or not) the null hypothesis of no heterogeneity. However, it has low power in meta-analyses if there are few primary studies and/or studies with small sample sizes, while exhibiting over-inflated power to detect small amounts of heterogeneity in meta-analyses with a large number of primary studies[2]. The I^2 value is frequently used to assess the relative extent of inconsistency. Nevertheless, it also has important limitations, as it is influenced by the sample size of the included primary studies (e.g., it may overestimate inconsistency across studies with precise estimates) and may yield biased results when used in the context of small-sample meta-analyses [1, 3, 4]. In addition, both the Q-Cochran test and the I^2 value are based on frequentist methods [5], potentially limiting their application to the Bayesian context [6].

These classical inconsistency measures exclusively rely on statistical criteria. However, there may be scenarios where concerns for apparently large statistical heterogeneity might be mitigated. In fact, the GRADE approach uses four items to judge inconsistency, namely the I^2, Cochran's Q test *p*-value, overlap in confidence intervals of primary studies by visual inspection, and the degree of difference in the point estimates of relative effects. To bring value to considered judgment, GRADE states that guideline and systematic review developers may abstain from rating down the certainty of evidence (CoE) if point estimates of primary studies are on the same side of a prespecified threshold (i.e., fall within the same target of the certainty range), despite the evidence of statistical heterogeneity [1, 7]. This assessment depends on providing context to outcomes' interpretation, by defining health outcome-level decision thresholds (DTs) – effect size measures suggesting whether an intervention translates to trivial or no, small, moderate or large effects [8, 9]. In a Bayesian context, it is possible to directly assess the proportion of effect sizes, sampled from the posterior distribution of the different primary studies, falling into the different ranges (interpretations) defined by DTs. Based on that, and on the concept of incorporating outcome-level DTs in inconsistency assessment, we developed two measures to support the assessment of inconsistency in meta-analysis. While the concept for this approach has been developed considering a Bayesian framework, it is also fully applicable to the frequentist context.

Given the limitations in existing approaches and their interpretation, our objective was to develop measures to support assessments of inconsistency. In this article we describe the development and application of two new measures : (i) one assessing overall outcome-level-related inconsistency (the Decision Inconsistency index), and (ii) one assessing across-studies inconsistency (the Across-Studies Inconsistency index). These measures are not intended to replace but rather to complement existing approaches to appraise inconsistency. We will start by reviewing the concept of the Dissimilarity Index

as the foundation from which our approach is derived. We will provide the concepts and formulae for the Decision Inconsistency Index and for the Across-Studies Inconsistency Index. Subsequently, we will apply our proposed approach using two practical examples. We will then present a web app and the metainc R package to implement the Decision Inconsistency Index and the Across-Studies Inconsistency Index. Finally, we will discuss potential limitations of our approach and how it may contribute to interpreting inconsistency in the GRADE CoE framework [10].

2. The Decision Inconsistency Index

2.1. The Dissimilarity Index

The Dissimilarity Index is one of the most commonly used demographics measures of segregation [11], reflecting the relative distributions of two groups over a set of geographic units [12]. It ranges between 0 and 1, with 0 indicating perfect integration (i.e., each geographic unit has the same percentage of members of each group as the total population) and 1 indicating maximum segregation (i.e., each geographic unit exclusively includes members of one of the two groups [11]) (Supplementary Figure 1). The formula for the computation of the Dissimilarity Index is the following [12, 13]:

$$Dissimilarity\ ind \quad = \frac{1}{2}\sum_{i}^{n}\left|\frac{N_{1i}}{N_1} - \frac{N_{2i}}{N_2}\right|$$

with n corresponding to the number of geographic units, N_{1i} corresponding to the population of group 1 in the geographical unit i, N_1 corresponding to the total population of group 1 in all considered geographical units, N_{2i} corresponding to the population of group 2 in the geographical unit i, and N_2 corresponding to the total population of group 2 in all considered geographical units being.

2.2. The Decision Inconsistency Index

Consider a meta-analysis including k primary studies comparing an intervention I *versus* a comparator C on a certain outcome whose reduction would correspond to a benefit and whose increase would correspond to a harm. For that outcome, an outcome-level DT has been established so that:

- If the effect size (ES) > DT, I would be associated with at least small harms.
- If ES < $-DT$, I would be associated with at least small benefits;
- If $-DT \leq$ ES $\leq DT$, I would be associated with a trivial or no effect (henceforth referred to as "trivial effect");

Therefore, in this scenario, we consider three interpretation categories (at least small benefits, trivial or no effects, and at least small harms) for ES, with two DT ($-DT$ and DT). We consider ES to be presented as absolute effects, as recommended by the GRADE working group for contextualizing ES in relation to DTs.

In a Bayesian meta-analytical context with no overall decision-related inconsistency, all samples from the posterior distributions of the ES of included primary studies will have their values associated with the same interpretation (either at least small benefits, at least small harms, or trivial effect). On the other extreme, if there is complete inconsistency, there will be a perfectly even distribution of the posterior ES samples across the three interpretation categories. That is, one third of the samples will indicate at least small benefits, another third will indicate at least small harms and the final third will indicate a trivial effect. Therefore, a situation of no clinical inconsistency would be analogous to one with full segregation in the geographical context (Dissimilarity Index=1), while a situation with maximum inconsistency would be analogous to one with full integration (Dissimilarity Index=0). However, in contrast with the Dissimilarity Index, we do not compare two groups but rather one distribution of posterior samples for the ES (across interpretation categories) with the expected distribution that would have been observed if there was maximum inconsistency. This concept forms the basis of a novel measure of inconsistency we propose – the Decision Inconsistency Index (DI). The DI quantifies overall inconsistency from a decision point of view, and may be calculated by:

$$DI = 1 - \left(\frac{\frac{1}{2}\sum_j^J \left| \frac{N_j}{N} - \frac{1}{J} \right|}{\frac{J-1}{J}} \right)$$

with N_j corresponding to the number of ES posterior samples per interpretation category, N corresponding to the total number of ES samples (i.e., the sum, for all primary studies, of all study-level posterior samples), and J corresponding to the number of interpretation categories.

Dividing by $\left(\frac{J-1}{J}\right)$ ensures that the DI lies between 0 and 1, while subtracting the ratio from 1 ensures that higher values are associated with higher inconsistency (as with the I^2 value). That is, the DI ranges between 0 and 1 (or 0-100%, if multiplied by 100), with 0 indicating no DT-related inconsistency and 1 indicating maximum DT-related inconsistency.

The DI can be calculated for as many interpretation categories as desired. If only two DTs are being considered (i.e., DT distinguishing trivial effects from at least small benefits and DT distinguishing trivial effects from at least small harms), three interpretation categories are possible (at least small benefits, at least small harms, and trivial effect), the formula of the DI can be stated as:

$$DI_{[2\ \mathrm{DTs}]} = 1 - \left(\frac{\frac{1}{2}\sum_j^3 \left| \frac{N_j}{N} - \frac{1}{3} \right|}{\frac{2}{3}} \right)$$

However, when using the GRADE Evidence to Decision (EtD) framework, decision-makers are usally interested in knowing not only whether an intervention is associated with non-trivial effects but also their magnitude (i.e., whether the interventions' desirable and undesirable health effects are trivial or none, small, moderate, or large) [14, 15]. The DI can be applied to these situations with three DTs on

each side of the no-effect (i.e., three DTs for benefits and three for harms) and, therefore, seven interpretation categories as recommended by GRADE [14, 15]. In this case, the formula of the DI can be stated as:

$$DI_{[6\ \mathrm{DTs}]} = 1 - \left(\frac{\frac{1}{2}\sum_{j}^{7} \left| \frac{N_j}{N} - \frac{1}{7} \right|}{\frac{6}{7}} \right)$$

3. The Across-Studies Inconsistency index

Although the DI provides a measure of the overall inconsistency of ES across interpretation categories, it does not quantify how inconsistent the ES of primary studies are when compared with each other. Let us consider the examples depicted in Figure 1. Figure 1A provides an example in which all ES samples of primary studies point to a trivial effect. The DI would be of 0%. Figure 1B and Figure 1C are two examples for which one third of ES samples suggest important benefits, one third points to important harms, and one third to a trivial effect. Therefore, in both examples, the DI would be expected to be large. However, while in Figure 1B all primary studies display a similar proportion of ES samples in each decision category, in Figure 1C the different primary studies display a different proportion of ES samples in each decision category (e.g., the first primary study would be expected to have most samples suggesting at least small harms while the third would be expected to have most samples suggesting at least small benefits). Therefore, the examples depicted in Figure 1B and 1C would differ on across-studies inconsistency, which would be larger in the latter.

Therefore, as a complement to the DI, we suggest that across-studies inconsistency should also be measured. For this purpose, Dissimilarity Index-based measures would not be suitable, as (i) the Dissimilarity Index has been devised to consider two groups, and (ii) it displays important limitations when dealing with small unit sizes (the Dissimilarity Index is extremely sensitive to small differences in cases when a small number of observations falls within a certain category)[16]. The adjustments proposed to address this limitation cannot be applied to measures generalizing the Dissimilarity Index to more than two groups[16, 17].

Therefore, to assess across-studies inconsistency, we propose a measure comparing (i) the observed number of samples per interpretation category for each study with (ii) the expected number of samples (per interpretation category for each study) if the proportion of samples per interpretation category had been the same for all studies and equal to the overall proportion. Potentially adequate candidates for such measures would be, for example, those based on the chi-squared statistic (χ^2), particularly relative to the maximum value (in order to allow for obtaining a measure ranging between 0 and 1). Given that, in this context, χ^2 would be given by:

$$\chi^2 = \sum \frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}}$$

(with N_{ij} corresponding to the number of ES samples per interpretation category and primary study, E_{ij} corresponding to the expected number of ES samples per interpretation category and primary study), and that the maximum of the chi-square statistic ($\max_{\chi^2}$) would be given by:

$$\max_{\chi^2} = N(\min(J, k) - 1),$$

(with J corresponding to the number of interpretation categories and k to the number of primary studies), then the Across-Studies Inconsistency Index (ASI) would be given by:

$$ASI = \sqrt{\frac{\sum \frac{\left(N_{ij} - E_{ij}\right)^2}{E_{ij}}}{N(\min(J, k) - 1)}}$$

The ASI measures across-study inconsistency considering decision interpretation categories. Its values can range between 0 and 1 (or 0-100%, if multiplied by 100), with 0 indicating no across-studies inconsistency and 1 indicating complete across-studies inconsistency.

4. Extension to the frequentist context

The DI and the ASI have been developed as measures using Bayesian meta-analysis. Indeed, the fact that Bayesian models yield a posterior distribution for the ES measures of primary studies renders that context particularly suited for the computation of the DI and ASI. However, the DI and ASI can also be calculated in the frequentist context. To accomplish that, one first needs to obtain the best linear unbiased predictions of the ES of each primary study, which can be calculated for frequentist random effects models empirically utilizing the obtained estimate for the between-study variability [τ^2] (which is analogous to the posterior ES estimates obtained in the Bayesian context). Subsequently, using these values and the corresponding standard-errors, it is possible to fit a probability distribution for each primary study, based on which samples can be drawn. The DI and ASI can then be calculated – in a similar way as in the Bayesian context – based on the sampled values for the best linear unbiased predictions of the ES of each primary study. Given that this is not straightforward, we developed an application for meta-analysts (see below).

5. Practical examples and application in sensitivity analyses

The Supplement displays two examples in which the DI and ASI are calculated. Supplementary Example 1 involves the computation of these indices in what was Cochrane's first living systematic review and meta-analysis of 18 primary studies comparing heparin with placebo on mortality at 12 months in a population of ambulatory patients with cancer [18] (Supplementary Tables 1-2). In brief,

this meta-analysis indicated that heparin was associated with decreased odds of mortality (5 fewer deaths per 1000 individuals; 95% credible interval: between 49 fewer deaths and 28 more deaths per 1000 individuals), with a I^2 value suggesting moderate inconsistency (I^2=35.2%). Computing the DT-based inconsistency indices, we would obtain a DI of 80.1% and a ASI of 19.3%, pointing to the possibility of relevant inconsistency (see below). The DI reflects the wide spread of posterior samples of ES across interpretation categories defined by information sizes, pointing to the potential difficulty in judging the size of the effect associated with the use of heparin.

We explored approaches for conducting sensitivity analyses, including (i) leave-one-out sensitivity analysis (removing each primary study at once and recalculating the DI and the ASI), (ii) sensitivity analysis based on risk of bias, (iii) sensitivity analysis based on uncertainty in baseline risk, and (iv) sensitivity analysis based on DTs. Overall, leave-one-out meta-analysis did not allow us to identify any individual primary study largely responsible for the observed inconsistency (Supplementary Table 3). The sensitivity analysis based on the risk of bias suggested that, in this example, inconsistency may be higher for studies displaying a lower risk of bias (DI=83.8%; ASI=23.2%) than for those with a high risk of bias (DI=77.9%; ASI=18.4%), but the difference was small. Finally, an increase in the baseline risk was associated with a decreasing trend for the DI (Supplementary Figure 2); the results of the sensitivity analysis based on DTs are displayed in Supplementary Figure 3.

Supplementary Example 2 computes the DI and ASI in a sample of 100 published meta-analyses of outcomes with available DTs [19, 20] (Supplementary Table 4). We observed that, in our sample, the median DI and ASI values were of 32% and 19%, respectively. On the other hand, the second tertile values are close to $DI \geq 50\%$ and $ASI \geq 25\%$ (Table 1). Considering only the DT of going from trivial or no to a small effect (sometimes consistent with the "minimal important difference") instead of three decision thresholds on each side of the null effect (trivial or no to small, small to moderate and moderate to large effects) did not produce a predictable effect on the DI (mean difference of 0.4 percent points) but was associated with an increase in the ASI (mean difference of 7.5 percent points) (Supplementary Figure 4).

6. Implementation in practice

We have developed an online app allowing for the computation of the DI and the ASI. The app, which is available at https://metainc.med.up.pt/ , takes as input a dataset containing the ES and the variance for each primary study. Based on the provided input, it can perform either frequentist meta-analysis using meta, or Bayesian meta-analysis using brms. Based on the posterior samples of the ES measures of the primary studies, the app provides information on the DI and on the ASI. Users of this online app are not required to (i) conduct a Bayesian meta-analysis or have knowledge on how to do it, or (ii) have knowledge on how to obtain the best linear unbiased predictions of the ES of primary studies in the

context of frequentist meta-analysis. Therefore, the app allows to overcome potential barriers for the computation of the DI and the ASI.

We have also developed a R package – metainc – to assess overall decision-related inconsistency and across-studies inconsistency (computing the DI and the ASI, respectively) after performing Bayesian or frequentist meta-analysis. It is available on CRAN (https://cran.r-project.org/web/packages/metainc/index.html). Supplementary Boxes 1-2 provide information and guided examples on how to use the metainc package.

7. Discussion

In this paper, we propose a quantitative approach to assess inconsistency in the meta-analytical context using DTs. This approach involves the computation of the DI and the ASI, which provide complementary information – the DI informs about the overall inconsistency of ES across interpretation categories, while the ASI quantifies across-studies inconsistency. The proposed measures have been developed in the Bayesian context, but they can also be computed for frequentist meta-analysis.

The GRADE guidance states that the assessment of inconsistency should not solely rely on classical measures of heterogeneity, as they have statistical limitations [1]. However, the current guidance for assessing inconsistency beyond those measures is centred on the visual inspection of the forest plot and plausibility of subgroup analyses. While our proposed methods are not intended to replace other approaches, they could provide valuable complementary information. In particular, by making use of DTs, the proposed methods can help interpreting the importance of observed inconsistency, something which is in line with recent statements to move away from interpreting results solely based on statistical criteria [21-23]. As an example, the inconsistency indices and their use of DTs can help identify situations in which (i) across-study inconsistency would impact the importance of findings (i.e., by providing a formal degree of contextualization), or (ii) statistical measures of heterogeneity may be overestimating inconsistency (e.g., due to primary studies with high precision estimates). Therefore, our proposed approach can be applied not only within GRADE, but also in the context of any meta-analysis, in order to help interpreting and framing the observed inconsistency.

Importantly, the proposed methods allow for sensitivity analyses based on the uncertainty in baseline risk or DTs. This accounts for potential uncertainty in the baseline risk or DTs when assessing inconsistency, which is not otherwise possible using only classical measures of heterogeneity or the visual inspection of the forest plot. Accounting for uncertainty may be particularly relevant since GRADE has proposed an approach to empirically obtain DTs that not only allows for the computation

of DT point estimates but also of best- and worst-case scenario DTs (Wiercioch et al, accepted pending revision, [9]). Performing sensitivity analyses based on DTs allows assessing whether inconsistency results are similar across a set of plausible DTs that can range between the best- and worst-case scenario DT values. In addition, based on sensitivity analyses, it is also possible to appraise unexplained inconsistency (i.e., inconsistency not explained by any pre-specified or convincing effect modifier) using the DI and the ASI. This is particularly relevant, since GRADE recommends that judgments on inconsistency are based on unexplained inconsistency.

Box 1 provides demonstrative examples (and a suggested reporting language) on how the DI and the ASI can be used to support inconsistency judgements in the assessment of the CoE in the GRADE approach. Although both the meta-analyses of the examples A and B display severe inconsistency as assessed by the I^2 value (example A: I^2=96%; example B: I^2=69%), the DT-related indices suggest that inconsistency may be at least a serious concern in example A (DI=73%; ASI=60%), but not in example B (DI=2%; ASI=9%; despite quantitative differences, the effect sizes are pointing to large or moderate benefits for all primary studies). While these examples illustrate a possible use of the DI and the ASI, this paper does intend to provide guidance on how to judge inconsistency in the GRADE approach. This is dealt with elsewhere including in the updated GRADE handbook (now called the GRADE Book https://book.gradepro.org/guideline/inconsistency [10]). It will require a broad agreement on a framework on how to consider different scenarios based on possible agreements and disagreements between the different items related to inconsistency (i.e., visual inspection of the forest plot, statistical measures of heterogeneity, and decision threshold-based inconsistency indices). This will also require definite guidance when it is adequate to downgrade inconsistency by two or even three levels for inconsistency. In addition, it will serve to adequately and jointly assess inconsistency and imprecision in order to avoid double penalisation. The DI and ASI could support answering such questions by providing information based on DTs in line with GRADE guidance (Wiercioch et al., accepted for publication pending revision).

The proposed approach has some limitations. Firstly, the assessment of inconsistency based on DTs should not be based solely on the calculation of the DI and the ASI. These measures should not be understood as definite indicators of inconsistency but as additional tools to use when assessing this domain. Another limitation concerns the absence of cut-off points defining low, moderate and severe inconsistency in the context of the DI and the ASI. We have evaluated the distribution of these indices in a sample of 100 meta-analyses, and this may provide suggestions to users on how to interpret their DI and ASI results. That is, users may hint at the magnitude of their inconsistency by comparison with the percentiles of DI and ASI for other systematic reviews. Nevertheless, setting of cut-off points for claiming inconsistency may require a more comprehensive approach and may depend on the number of considered decision thresholds. However, it should be noted that universally-agreed or even sensible cut-off points do not exist for the I^2 either [2]. Exploration of other approaches to explore inconsistency

ratings may be useful and we are beginning work in a GRADE project group to provide this guidance. Presenting only the proportions point estimates falling into within different certainty ranges may be one alternative. However, it brings complexity of having to interpret many data points simultaneously, requiring judgments with unknown reliability (particularly in meta-analyses with a small amount of primary studies) and not considering the confidence intervals of the studies' estimates. Finally, not all functions are currently available for non-R users. However, efforts are being made to increase the number of functions accessible through different software or platforms.

8. Conclusion

In meta-analysis, the assessment of inconsistency based solely on classical measures of heterogeneity has important limitations. Considering DTs may allow for that assessment in the respective health decision context. However, no quantitative approaches had been proposed so far. In this paper, we describe two measures – the DI and the ASI – that can be used to quantitatively assess inconsistency using DTs. While their computation does not replace other methods for assessing inconsistency, used together they can be particularly helpful for (i) interpreting the health importance of observed inconsistency, (ii) giving the evaluator additional information about the impact of potential uncertainty in baseline risk or DTs, and (iii) supporting the rating of inconsistency in the GRADE appraisal of the CoE. Based on the developed R package and web tool, this approach can be easily implemented both in the Bayesian and frequentist contexts.

Acknowledgements: The authors would like to thank the Evidence Synthesis Hackathon, whose 2023 edition allowed for relevant advances on this study and on the respective R package.

Author contribution statement:

- BSP has participated in conceptualization, data curation, formal analysis, methodology, and writing - original draft;
- MMC and SGM have participated in formal analysis, and writing – review & editing;
- SM, CN, GB and PW have participated in methodology, and writing – review & editing;
- GuS, JMS and GaS have participated in data curation, formal analysis, methodology, and writing – review & editing;
- IN, RJV, AB, HJS and LFA have participated in conceptualization, methodology, and writing - review & editing.

Declarations of interest: HJS is co-chair of the GRADE Working Group, but this is not an official GRADE Working Group article (although the concepts herein may be used by GRADE in the future but this will require formal approval). All other authors declare no conflict.

Funding sources: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Generative AI and AI-assisted technologies in the writing process: No generative AI or AI-assisted technologies were used in the writing process or in any other task for this manuscript.

References

[1] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64:1294-302.

[2] Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. Cochrane handbook for systematic reviews of interventions. 2019:241-84.

[3] von Hippel PT. The heterogeneity statistic I(2) can be biased in small meta-analyses. BMC Med Res Methodol. 2015;15:35.

[4] Schunemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. J Clin Epidemiol. 2022;150:225-42.

[5] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-58.

[6] Welton NJ, Sutton AJ, Cooper N, Abrams KR, Ades A. Evidence synthesis for decision making in healthcare: John Wiley & Sons; 2012.

[7] Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. J Clin Epidemiol. 2023;158:70-83.

[8] Morgano GP, Mbuagbaw L, Santesso N, Xie F, Brozek JL, Siebert U, et al. Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomised methodological study (GRADE-THRESHOLD). BMJ Open. 2022;12:e053246.

[9] Morgano GP, Wiercioch W, Piovani D, Neumann I, Nieuwlaat R, Piggott T, et al. Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a randomised methodological study (GRADE-THRESHOLD). J Clin Epidemiol. 2024:111639.

[10] Neumann I, Sousa-Pinto B, Meerpohl J, Dahm P, Brennan S, Alonso-Coello P, et al. Inconsistency. In: Neumann I, Schünemann H, editors. GRADE Book. https://book.gradepro.org: GRADE Working Group; 2024.

[11] Massey DS, Rothwell J, Domina T. The Changing Bases of Segregation in the United States. Ann Am Acad Pol Soc Sci. 2009;626.

[12] Massey DS, Denton NA. The dimensions of residential segregation. Social forces. 1988;67:281-315.

[13] The Social Science Data Analysis Network. CensusScope: ABOUT DISSIMILARITY INDICES.

[14] Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016.

[15] Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089.

[16] Allen R, Burgess S, Davidson R, Windmeijer F. More reliable inference for the dissimilarity index of segregation. Econom J. 2015;18:40-66.

[17] Sakoda JM. A generalized index of dissimilarity. Demography. 1981;18:245-50.

[18] Akl EA, Kahale LA, Hakoum MB, Matar CF, Sperati F, Barba M, et al. Parenteral anticoagulation in ambulatory patients with cancer. Cochrane Database Syst Rev. 2017;9:CD006652.

[19] Cuker A, Tseng EK, Nieuwlaat R, Angchaisuksiri P, Blair C, Dane K, et al. American Society of Hematology living guidelines on the use of anticoagulation for thromboprophylaxis in patients with COVID-19: January 2022 update on the use of therapeutic-intensity anticoagulation in acutely ill patients. Blood Adv. 2022;6:4915-23.

[20] Piggott T, Nonino F, Baldin E, Filippini G, Rijke N, Schunemann H, et al. Multiple Sclerosis International Federation guideline methodology for off-label treatments for multiple sclerosis. Mult Scler J Exp Transl Clin. 2021;7:20552173211051855.

[21] Hernán MA, Greenland S. Why Stating Hypotheses in Grant Applications Is Unnecessary. JAMA. 2024.
[22] Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Taylor & Francis; 2016. p. 129-33.
[23] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond “p< 0.05”. Taylor & Francis; 2019. p. 1-19.

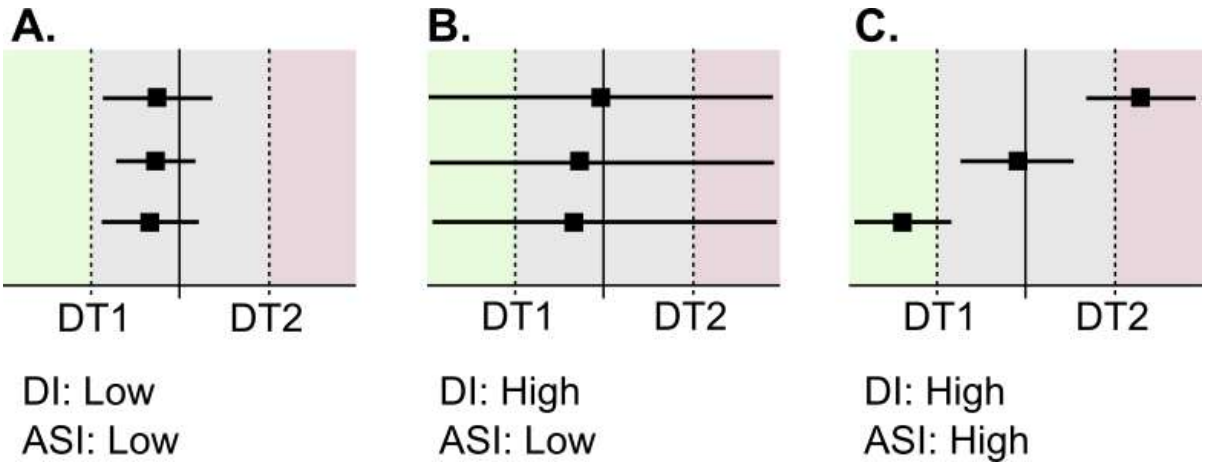
Tables

Table 1. Summary of the distributions of the Decision Inconsistency Index (DI) and Across-Studies Inconsistency Index (ASI) in a sample of 100 published meta-analysis

Inconsistency measure	Percentile 50 (median)	Percentile 67	Percentile 75	Maximum
DI (%)	32.2	49.1	58.3	86.2
ASI (%)	19.2	24.8	28.6	60.5

Figures

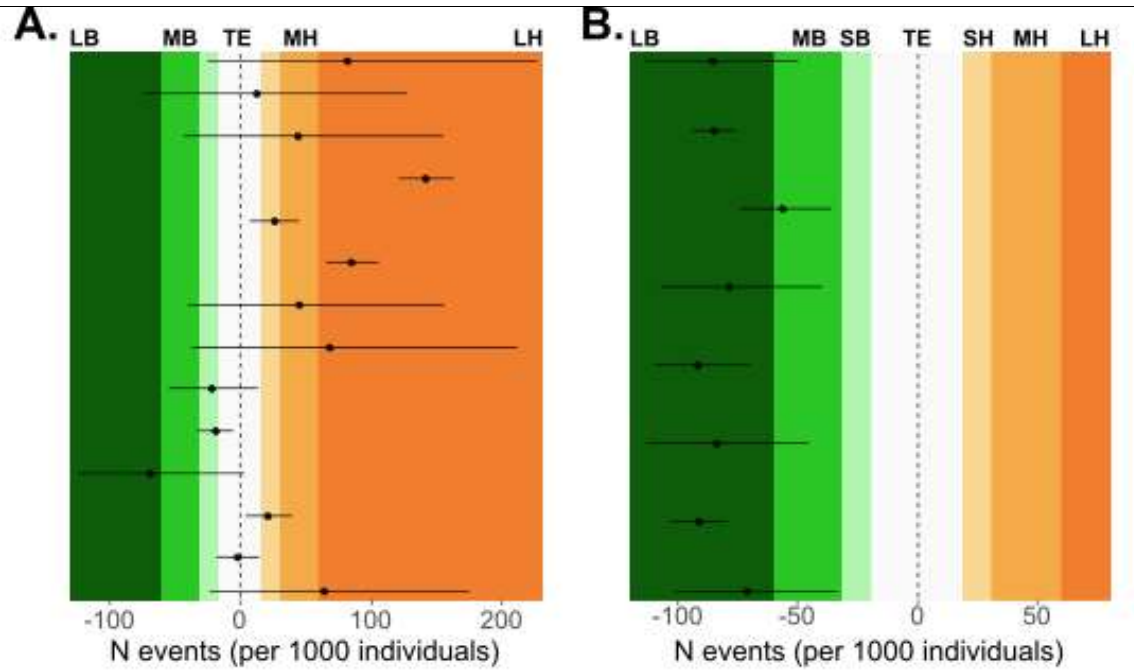
Figure 1. Hypothetical meta-analytical examples illustrating the differences in the concepts of Decision Inconsistency (DI) and Across-Studies Inconsistency (ASI). The green zone (left) indicates at least small benefits, the grey zone (centre) indicates a trivial or no effect, and the red zone (right) indicates at least small harms.



In Figure 1B, even though the point estimates of primary studies would all indicate a trivial effect, the DI would be high as there would be a large proportion of posterior samples also indicating at least small benefits and at least small harms (the effects of the primary studies are all compatible with at least small benefits, a trivial effect and at least small harms); however, all studies would be similar among themselves (hence, the ASI would be low). In Figure 1C, both the DI and the ASI would be high as (i) there would be a large proportion of posterior samples indicating at least small benefiis, a trivial effect or at least small harms, and (ii) all studies would be very different among themselves. DT=Decision thresholds

Boxes

Box 1. Examples on how the Decision Inconsistency Index (*DI*) and the Across-Studies Inconsistency (*ASI*) indices can be used to support judgements on inconsistency in GRADE



Example A: We had <u>at least serious</u> concerns about inconsistency of the evidence. High inconsistency was suggested both by statistical measures of heterogeneity (I^2 = 96%) and by threshold-based inconsistency indices (*DI* = 73%; *ASI* = 60%). No variable was identified that would potentially explain the inconsistency. We therefore rated down the certainty of evidence for inconsistency by at least 1 level.

Example B: We had <u>no serious</u> concerns about inconsistency of the evidence. Although high inconsistency was suggested by statistical measures of heterogeneity (I^2 = 69%), decision threshold-based inconsistency indices suggested low inconsistency (*DI* = 2%; *ASI* = 9%). This shows that inconsistency may not be important. We therefore rated down the certainty of evidence for inconsistency by 0 levels.

Declarations of interest: HJS is co-chair of the GRADE Working Group, but this is not an official GRADE Working Group article (although the concepts herein may be used by GRADE in the future but this will require formal approval). All other authors declare no conflict.

Author contribution statement:

- BSP has participated in conceptualization, data curation, formal analysis, methodology, and writing - original draft;
- MMC and SGM have participated in formal analysis, and writing – review & editing;
- SM, CN, GB and PW have participated in methodology, and writing – review & editing;
- GuS, JMS and GaS have participated in data curation, formal analysis, methodology, and writing – review & editing;
- IN, RJV, AB, HJS and LFA have participated in conceptualization, methodology, and writing - review & editing.