Facilitating GRADE judgements about the inconsistency of effects using a novel visualisation approach

Mohammad Hassan Murad ,[1] Zhen Wang ,[1,2] Yngve Falck-Ytter[3]

10.1136/bmjebm-2024-113038

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjebm-2024-113038).

[1]Evidence-based Practice Center, Mayo Clinic, Rochester, MN, USA
[2]Health Care Policy and Research, Mayo Clinic Minnesota, Rochester, Minnesota, USA
[3]Case Western Reserve University, Cleveland, Ohio, USA

Correspondence to: **Dr Mohammad Hassan Murad;** murad.mohammad@mayo.edu

Check for updates

To cite: Murad MH, Wang Z, Falck-Ytter Y. *BMJ Evidence-Based Medicine* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjebm-2024-113038

Background

Inconsistency is a key domain that determines the certainty of evidence. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach specifically defines inconsistency as the variability in results across studies, and not variability in study characteristics, eligibility criteria or design.[1] Statistical measures of heterogeneity are often used to assess inconsistency, however, major limitations of such measures have been described. For example, Cochran's Q test for homogeneity is usually underpowered to detect heterogeneity. The I^2 index which is the most commonly used measure, underestimate true statistical heterogeneity when there are fewer than 10 studies in a meta-analysis, which is a common scenario, and is correlated with the sample size of the included studies.[2] The I^2 index is also often misunderstood as an indicator of the spread of the effect size. Borenstein demonstrates how a meta-analysis with I^2 index of 25% can have more spread of the effect size than a meta-analysis with I^2 index of 75%.[3] Therefore, GRADE guidance on inconsistency recommended less reliance on statistical measures and instead, instructed to make judgements about whether studies in a meta-analysis provide estimates that are clinically importantly different from each other.[1] However, there are no existing tools to facilitate this process making it highly subjective. Users are instructed to look at a forest plot and evaluate the similarity of point estimates of the included studies and the overlap of their CIs, and make a judgement based on values that they consider clinically important. Merely counting studies does not work because some studies can be outliers but may have a very small weight within the pooled effect estimate. Having multiple thresholds makes this task even more difficult. Furthermore, in the case of binary outcomes, decision thresholds are based on absolute treatment effects[4 5] whereas most meta-analyses and their associated forest plot are performed on relative effect scales.

In this exposition, we operationalise the GRADE definition of decision thresholds (trivial, small, medium or large effect)[4] and judge inconsistency in a meta-analysis based on these thresholds. We developed a tool for visualising inconsistency based on stakeholder-provided thresholds. The first aim of this visualisation approach is educational, that is, to teach the concept of inconsistency as it relates to multiple decisional thresholds. The second aim of this visualisation is to provide a practical tool to facilitate making judgements about inconsistency in a meta-analysis or a guideline.

The proposed visualisation approach

This visualisation approach can be used when meta-analysts prepare a summary of the findings table and make a judgement about inconsistency. The approach starts with stakeholders providing three thresholds in the form of absolute risk differences. Consistent with recent GRADE guidance,[4] these three thresholds define seven treatment effect ranges consistent with large, medium and small reduction, large, medium and small increase, and a trivial or no effect. As an example, we used in this paper the following thresholds (per 1000 patients): small, medium and large reduction (−10 to −100 and −200), small, medium and large increase (10, 100, 200), and trivial or no effect (between −10 and +10). Random-effects meta-analysis is conducted using the restricted maximum likelihood estimator of between-study heterogeneity on a relative effect scale. The relative treatment effect of each study is converted to an absolute effect using a baseline risk that is either derived from the available studies or can also be provided by users (in this visualisation, it was derived by dividing the number of events in the control groups of a meta-analysis by the total number of participants in the control arms). Each individual study is categorised into one of the seven ranges based on its absolute effect. The random-effect weights of studies that fall in each inference range are summed to provide the total weight for that range. A bar graph depicts all the ranges with the height of the bars representing the percentage of the total weight for the range. This bar graph allows visualisation of the distribution of inferences of the individual studies in relation to stakeholder-provided thresholds. The approach is summarised in box 1. This approach can also be used for continuous outcomes, which can be expressed on their original scale and using stakeholder-provided thresholds. If such thresholds were unknown, the outcome could be expressed as a standardised mean difference and we can use the traditional thresholds of 0.2, 0.5 and 0.8 to define small, moderate and large effect thresholds.[4]

The approach is implemented for binary and continuous outcomes in an open-source R code

Box 1 Steps of performing the proposed visualisation approach

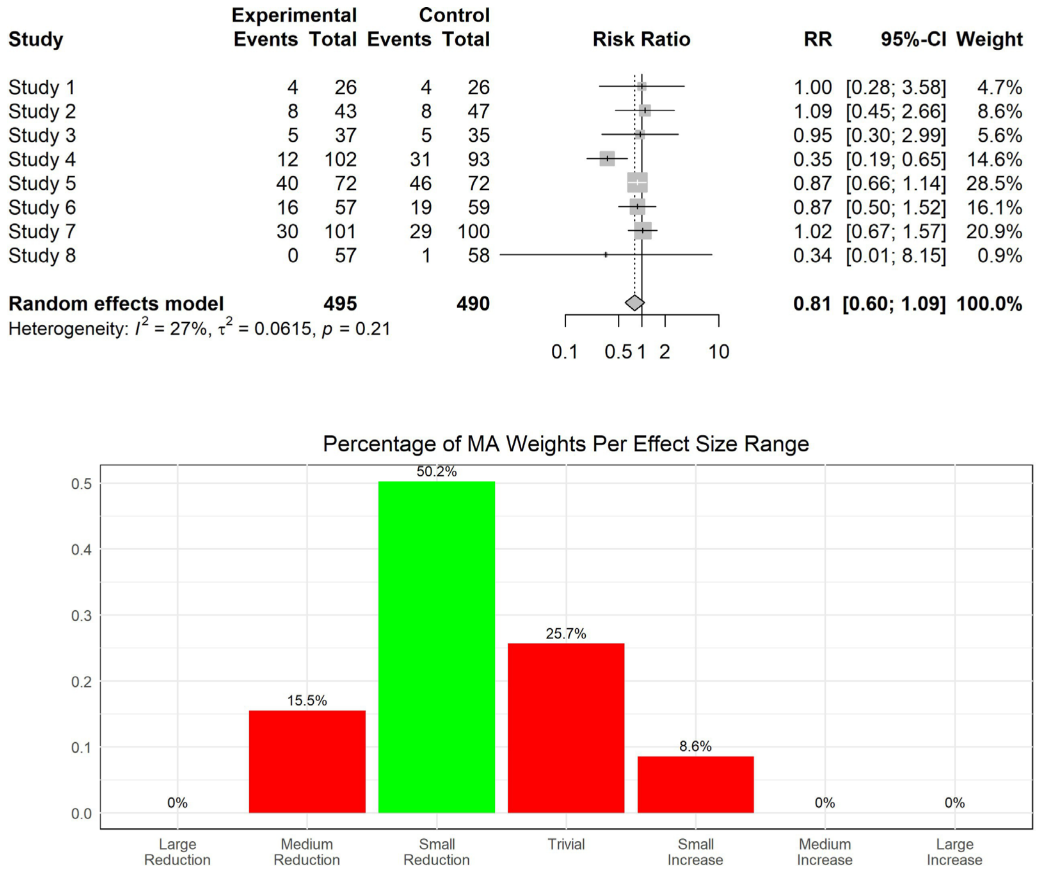
1. Decisional thresholds are provided by stakeholders.
   ⇒ If unavailable or unknown, default thresholds can be used.
2. Meta-analysis is conducted to obtain the effect size and weight for each study.
3. The effect size of each study is converted as needed to match units of decisional thresholds.
   ⇒ Binary outcomes: convert relative effects to absolute effects using appropriate baseline risk.
   ⇒ Continuous outcomes: convert to a standardised mean difference if thresholds are unavailable.
4. Total weight is calculated for each decisional range by summing the weights of individual studies that fall within that range.
5. A bar graph allows users to visualise the spread of inference across decisional ranges and make a judgement about inconsistency.

provided in the online supplemental appendix (R Core Team 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria). The code is implemented in an R Shiny application that does not require knowledge of statistical software coding: https://hassan-murad.shinyapps.io/inconsistency_visualization.

Examples

The first example is a meta-analysis of eight studies[6] that evaluated the effect of home non-invasive pressure ventilation on mortality in patients with chronic obstructive pulmonary disease. The I^2 index of 27% and the p value for heterogeneity of 0.21 suggest no important heterogeneity. The point estimates on the relative risk scale are similar except for two smaller studies (figure 1, panel A). Applying the proposed visualisation approach (figure 1, panel B), we note that the pooled effect and 50.2% of the weights of individual studies suggest a small reduction in risk. However, the remaining 49.8% of the weights of individual studies suggest three different inferences: moderate reduction, trivial effect and small increase. Thus, inferences from individual studies are quite variable in contrast to the impression derived from the forest plot and its associated statistical measures. In this case, rating down for inconsistency seems justified.

The second example is a meta-analysis of nine studies[7] that evaluated the adverse events of fluoxetine in patients who are obese or overweight. The I^2 index of 53% and the p value for heterogeneity of 0.03, suggest a substantial and statistically significant heterogeneity (figure 2, panel A). Applying the proposed visualisation approach (figure 2, panel B), suggests that the inference from almost all the studies (96% of the meta-analysis weight) is consistent with a small increase in



Study	Experimental Events	Experimental Total	Control Events	Control Total	RR	95%-CI	Weight
Study 1	4	26	4	26	1.00	[0.28; 3.58]	4.7%
Study 2	8	43	8	47	1.09	[0.45; 2.66]	8.6%
Study 3	5	37	5	35	0.95	[0.30; 2.99]	5.6%
Study 4	12	102	31	93	0.35	[0.19; 0.65]	14.6%
Study 5	40	72	46	72	0.87	[0.66; 1.14]	28.5%
Study 6	16	57	19	59	0.87	[0.50; 1.52]	16.1%
Study 7	30	101	29	100	1.02	[0.67; 1.57]	20.9%
Study 8	0	57	1	58	0.34	[0.01; 8.15]	0.9%
Random effects model		**495**		**490**	**0.81**	**[0.60; 1.09]**	**100.0%**

Heterogeneity: $I^2 = 27\%$, $\tau^2 = 0.0615$, $p = 0.21$

Figure 1 Meta-analysis of trials of home non-invasive positive pressure ventilation in chronic obstructive pulmonary disease. Panel A (top) demonstrates a meta-analysis of the risk ratio scale suggesting small heterogeneity. Panel B (bottom) demonstrates a bar graph showing the distribution of meta-analysis weights per effect size range, suggesting important inconsistency. The green bar represents the inference associated with the target of certainty (the pooled estimate). MA, meta-analysis.

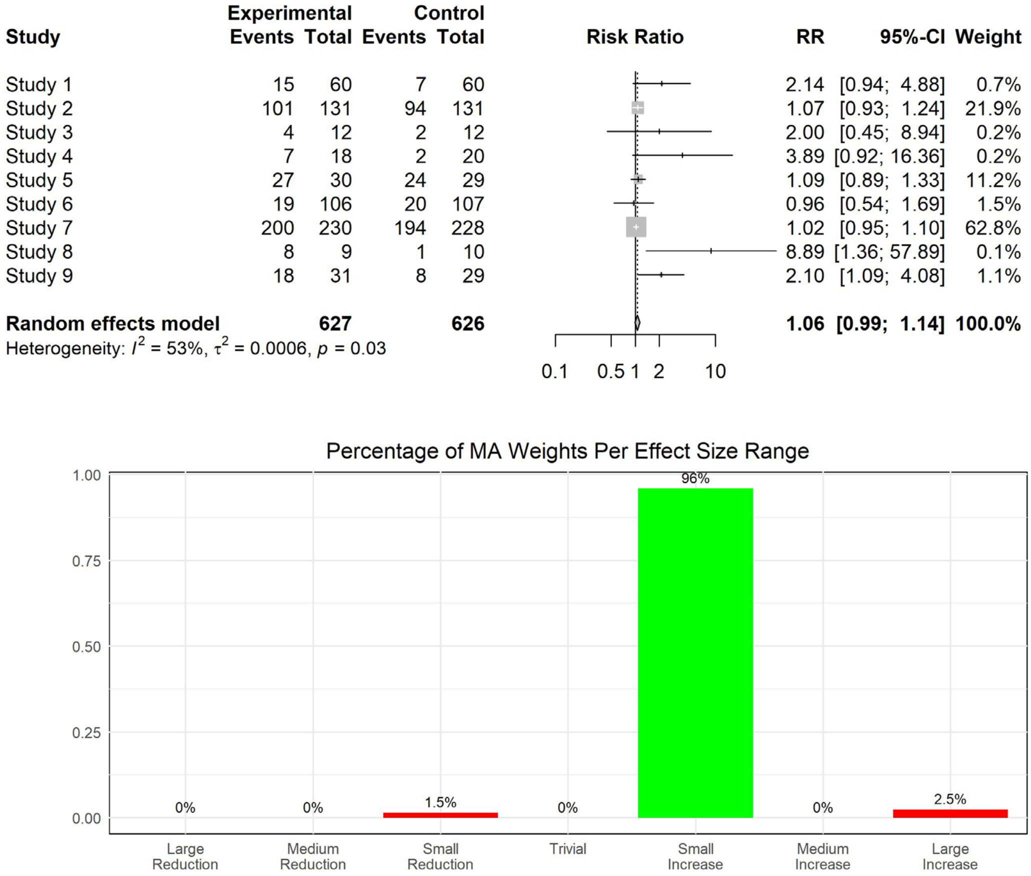


Figure 2 Meta-analysis of trials of fluoxetine for adults who are overweight or obese. Panel A (top) demonstrates a meta-analysis of the risk ratio scale suggesting substantial heterogeneity. Panel B (bottom) demonstrates a bar graph showing the distribution of meta-analysis weights per effect size range, suggesting minimal inconsistency. The green bar represents the inference associated with the target of certainty (the pooled estimate). MA, meta-analysis.

risk. Therefore, the statistically significant heterogeneity did not lead to any important inconsistency when considering stakeholder-provided thresholds. In this case, rating down for inconsistency is unnecessary.

The third example addresses a continuous outcome (online supplemental figure 1). A meta-analysis of eight trials evaluated the effect of health and wellness coaching on the severity of depression in patients with chronic illness.[8] The I^2 index of 95% and the p value for heterogeneity of 0.01 suggested a substantial and statistically significant heterogeneity. The proposed visualisation shows that the majority of evidence (75% of meta-analysis weight) was consistent with a single inference, a trivial effect, which can justify not rating down for inconsistency. The forest plot and the bar graph (online supplemental figure 1) demonstrate that statistical heterogeneity is driven by a single small study (11.9% of the weight). Reviewing the inclusion criteria for this study may indicate a systematic difference from the remaining eight studies.

Discussion

It is very challenging to look at a forest plot and judge the consistency of individual studies in terms of their inference relating to multiple inference regions, up to seven regions according to recent GRADE guidance.[4] This complexity increases to another level in the case of binary outcomes, which require translation to absolute effects. The proposed visualisation and quantification of total weight across stakeholder-provided thresholds can help in streamlining this judgement and make it more explicit. Using meta-analysis weights instead of 'counting studies' addresses small studies that are outliers with extreme results.

The approach can also be used when stakeholders decide to not to use multiple thresholds, and opt to only use the minimally important difference (MID). A positive MID and a negative MID define three ranges of effect, important reduction, trivial to no effect and important increase.[5] The same visualisation can show the total meta-analysis weight distributed across these three ranges to make a judgement about inconsistency.

Limitations to this approach include two concerns associated with transforming a relative effect of a binary outcome to an absolute one. The first issue is the assumption of portability of the relative effect across different baseline risks, which is not always true.[9] The second issue is that such transformation is usually done without addressing uncertainty in baseline risks. Several methods have been proposed to address uncertainty in the baseline risk when estimating the absolute effect,[10] which can be easily incorporated in this proposed visualisation approach. Lastly, there are inherent methodological limitations to the MID and its reliability for gauging clinical relevance. Other approaches for establishing clinical relevance thresholds exists.[11]

X Mohammad Hassan Murad @m_hassan_murad

Contributors MHM and YF-Y conceived this study. MHM wrote the first draft. All authors critically revised the manuscript and approved the final version.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Mohammad Hassan Murad http://orcid.org/0000-0001-5502-5975
Zhen Wang http://orcid.org/0000-0002-9368-6149

References

1 **Guyatt G**, Zhao Y, Mayer M, *et al*. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. *J Clin Epidemiol* 2023;158:70–83.

2 **Morton SC**, Murad MH, O'Connor E, *et al*. Quantitative synthesis-an update. In: *Methods guide for effectiveness and comparative effectiveness Reviews*. Rockville (MD), 2008.

3 **Borenstein M**. Research Note: In a meta-analysis, the I^2 index does not tell us how much the effect size varies across studies. *J Physiother* 2020;66:135–9.

4 **Schünemann HJ**, Neumann I, Hultcrantz M, *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 **Zeng L**, Brignardello-Petersen R, Hultcrantz M, *et al*. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. *J Clin Epidemiol* 2022;150:216–24.

6 **Wilson ME**, Dobler CC, Morrow AS, *et al*. Association of Home Noninvasive Positive Pressure Ventilation With Clinical Outcomes in Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-analysis. *JAMA* 2020;323:455–65.

7 **Serralde-Zúñiga AE**, Gonzalez Garay AG, Rodríguez-Carmona Y, *et al*. Fluoxetine for adults who are overweight or obese. *Cochrane Database Syst Rev* 2019;10:CD011688.

8 **Boehmer KR**, Álvarez-Villalobos NA, Barakat S, *et al*. The impact of health and wellness coaching on patient-important outcomes in chronic illness care: A systematic review and meta-analysis. *Pat Educ Couns* 2023;117:107975.

9 **Murad MH**, Chu H, Wang Z, *et al*. Hierarchical models that address measurement error are needed to evaluate the correlation between treatment effect and control group event rate. *J Clin Epidemiol* 2024;170:111327.

10 **Murad MH**, Wang Z, Zhu Y, *et al*. Methods for deriving risk difference (absolute risk reduction) from a meta-analysis. *BMJ* 2023;381:e073141.

11 **Dekker J**, de Boer M, Ostelo R. Minimal important change and difference in health outcome: An overview of approaches, concepts, and methods. *Osteoarthr Cartil* 2024;32:8–17.

Appendix

1. R code that generates inconsistency visualization for binary outcomes

Input: Studies should be in rows. Four columns are needed for the number of events (r1 and r2) and sample sizes (n1 and n2) for the intervention and control group, respectively.

```
library(meta); library(ggplot2)

thresholds <- c(-Inf, -0.2, -0.1, -0.01, 0.01, 0.1, 0.2, Inf)
labels <- c("Large\nReduction", "Medium\nReduction", "Small\nReduction",
       "Trivial", "Small\nIncrease", "Medium\nIncrease", "Large\nIncrease")

meta_object <- metabin(event.e = r1, n.e = n1, event.c = r2, n.c = n2,
              studlab=Study, data = dat, sm = "RR")
  BR <- sum(summary(meta_object)$event.c) / sum(summary(meta_object)$n.c)
  pooled.RD <- (exp(summary(meta_object)$TE.random) - 1) * BR
  pooled.reg <- labels[findInterval(pooled.RD, thresholds)]
  RDs <- ((exp(summary(meta_object)$TE)) - 1) * BR
  RDs_regions <- labels[findInterval(RDs, thresholds)]
  RDs_regions <- factor(RDs_regions, levels = labels)
  wts <- summary(meta_object)$w.random
  total_wt <- sum(wts)
  region_total_wt <- tapply(wts, RDs_regions, sum)
  region_total_wt[is.na(region_total_wt)] <- 0
  region_percent_wts <- region_total_wt / total_wt

  df_plot <- data.frame(labels = factor(labels, levels = labels),
                        region_percent_wts = region_percent_wts,
                        pooled_reg = pooled.reg)
ggplot(df_plot, aes(x = labels, y = region_percent_wts)) +
  geom_bar(stat = "identity", fill = ifelse(df_plot$labels == df_plot$pooled_reg, "green", "red")) +
  geom_text(aes(label = paste0(round(region_percent_wts * 100, 1), "%")),
            vjust = -0.5, size = 3) +
  labs(x = NULL, y = NULL) +
  ggtitle("Percentage of MA Weights Per Effect Size Range") +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "white"),
    plot.title = element_text(hjust = 0.5)) +
  theme(plot.margin = margin(1, 1, 1, 4, "lines"))
```

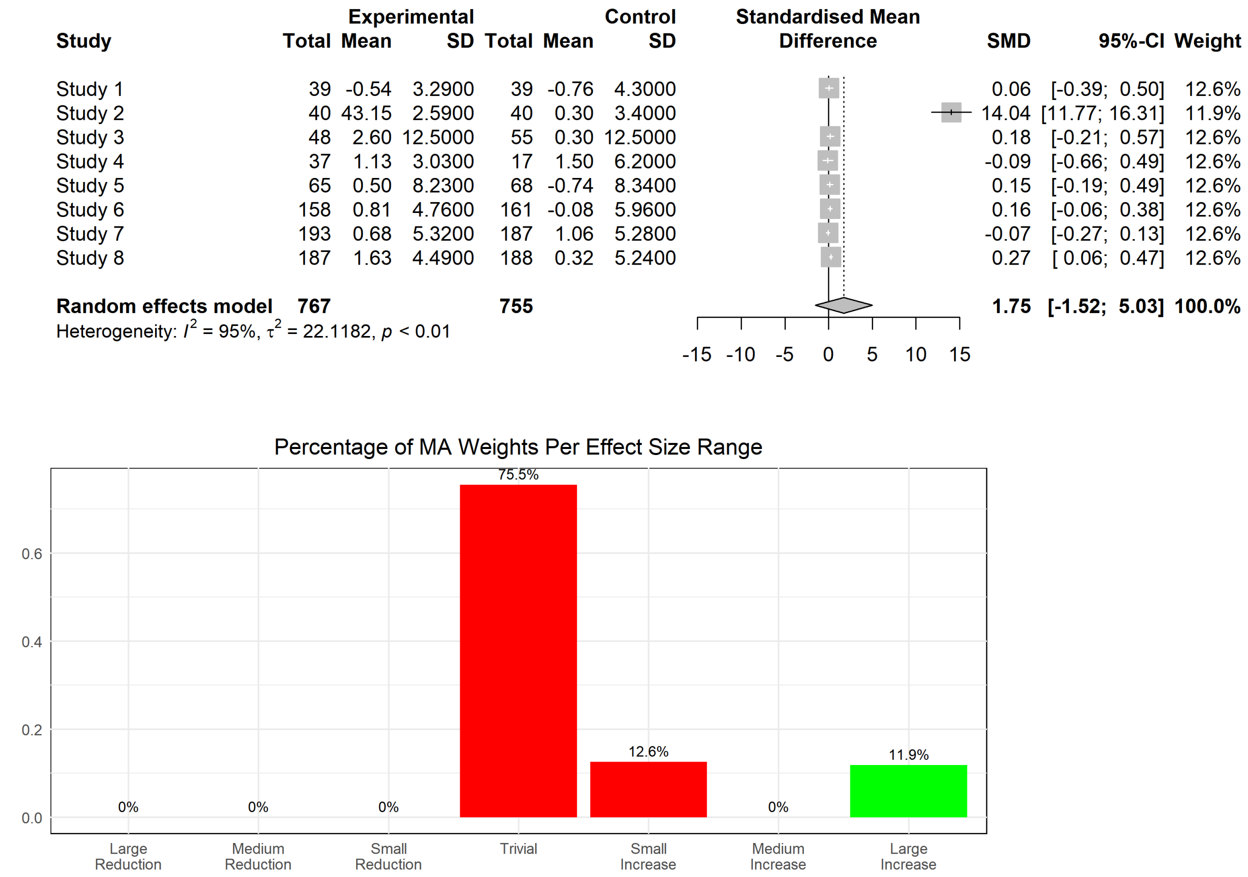
```
forest(meta_object, common = F)
```

2. R code that generates inconsistency visualization for continuous outcomes

Input: Studies should be in rows. Six columns are needed for sample sizes (n1 and n2), means (m1 and m2) and standard deviations (sd1 and sd2) for the intervention and control group, respectively.

```
library(meta); library(ggplot2)

thresholds <- c(-Inf, -0.8, -0.5, -0.2, 0.2, 0.5, 0.8, Inf)
labels <- c("Large\nReduction", "Medium\nReduction", "Small\nReduction",
       "Trivial", "Small\nIncrease", "Medium\nIncrease", "Large\nIncrease")

meta_object <- metacont(n.e = n1, mean.e = m1, sd.e=sd1, n.c=n2,mean.c=m2, sd.c=sd2,
              studlab=Study, data = dat, sm = "SMD")
  pooled.smd <- summary(meta_object)$TE.random
  pooled.reg <- labels[findInterval(pooled.smd, thresholds)]
  SMDs <- summary(meta_object)$TE
  SMDs_regions <- labels[findInterval(SMDs, thresholds)]
  SMDs_regions <- factor(SMDs_regions, levels = labels)
  wts <- summary(meta_object)$w.random
  total_wt <- sum(wts)
  region_total_wt <- tapply(wts, SMDs_regions, sum)
  region_total_wt[is.na(region_total_wt)] <- 0
  region_percent_wts <- region_total_wt / total_wt

  df_plot <- data.frame(labels = factor(labels, levels = labels),
                        region_percent_wts = region_percent_wts,
                        pooled_reg = pooled.reg)

ggplot(df_plot, aes(x = labels, y = region_percent_wts)) +
  geom_bar(stat = "identity", fill = ifelse(df_plot$labels == df_plot$pooled_reg, "green", "red")) +
  geom_text(aes(label = paste0(round(region_percent_wts * 100, 1), "%")),
            vjust = -0.5, size = 3) +
  labs(x = NULL, y = NULL) +
  ggtitle("Percentage of MA Weights Per Effect Size Range") +
  theme_minimal() +
  theme(
    panel.background = element_rect(fill = "white"),
    plot.title = element_text(hjust = 0.5)) +
  theme(plot.margin = margin(1, 1, 1, 4, "lines"))

png("fig3_panelA.png", width=12*ppi, height=4*ppi, res=ppi)
forest(meta_object, common = F)
```

Supplemental Figure 1. Meta-analysis of trials of health coaching for chronic illness



Study	Experimental Total	Experimental Mean	Experimental SD	Control Total	Control Mean	Control SD	SMD	95%-CI	Weight
Study 1	39	-0.54	3.2900	39	-0.76	4.3000	0.06	[-0.39; 0.50]	12.6%
Study 2	40	43.15	2.5900	40	0.30	3.4000	14.04	[11.77; 16.31]	11.9%
Study 3	48	2.60	12.5000	55	0.30	12.5000	0.18	[-0.21; 0.57]	12.6%
Study 4	37	1.13	3.0300	17	1.50	6.2000	-0.09	[-0.66; 0.49]	12.6%
Study 5	65	0.50	8.2300	68	-0.74	8.3400	0.15	[-0.19; 0.49]	12.6%
Study 6	158	0.81	4.7600	161	-0.08	5.9600	0.16	[-0.06; 0.38]	12.6%
Study 7	193	0.68	5.3200	187	1.06	5.2800	-0.07	[-0.27; 0.13]	12.6%
Study 8	187	1.63	4.4900	188	0.32	5.2400	0.27	[0.06; 0.47]	12.6%
Random effects model	**767**			**755**			**1.75**	**[-1.52; 5.03]**	**100.0%**

Heterogeneity: $I^2 = 95\%$, $\tau^2 = 22.1182$, $p < 0.01$

Bottom Panel B demonstrates a bar graph showing the distribution of meta-analysis weights per effect size range. The green bar represents the inference associated with the target of certainty (the pooled estimate).