# The choice and definition of summary measure for meta-analysis of clinical studies with binary outcomes: Effect on clinical interpretation

**Authors and institutions:**

Andrew A Plumb[1]

Steve Halligan[1]

Susan Mallett[2]

[1]Centre for Medical Imaging, University College London, 43-45 Foley St, London, W1W 7TS

[2]Institute of Applied Health Research, University of Birmingham, B15 2TT

**Correspondence to:**

Andrew Plumb, Associate Professor of Medical Imaging, Centre for Medical Imaging, University College London, 43-45 Foley St, London, W1W 7TS.

andrew.plumb@ucl.ac.uk

**Article type:**

Commentary

**Word Count (excluding figures and legends):**

2,287

**Figures:**

4

**Tables:**

1

# The choice and definition of summary measure for meta-analysis of clinical studies with binary outcomes: Effect on clinical interpretation

**Article type:**

Commentary

**Word Count (excluding figures and legends):**

2,287

**Figures:**

4

**Tables:**

1

**Abstract:**

Many systematic reviews and meta-analyses concern the effect of a healthcare intervention on a binary outcome i.e. occurrence (or not) of a particular event. Usually, the overall effect, pooled across all studies included in the meta-analysis, is summarised using the odds ratio (OR) or the relative risk (RR). Under most circumstances, it is obvious how to identify what should be considered as the event of interest – for example, death or a clinically-important side effect. However, on occasion it may not be clear in which "direction" the event should be specified – such as attendance (vs. non-attendance) at cancer screening. Usually, this choice is not critical to the overall conclusion of the meta-analysis, but occasionally it can lead to differences in how the included studies are pooled, ultimately affecting the overall meta-analytic result, particularly when using relative risks rather than odds ratios. In this commentary, we will explain this phenomenon in more detail using examples from the literature, and explore how analysts and readers can avoid some potential pitfalls.

**Commentary:**

There are a number of interesting examples in the literature whereby the choice of summary measure for a meta-analysis can substantially affect its clinical interpretation[1], sometimes leading the reader to draw different conclusions regarding the statistical significance of the results. A recently published article by Zhu et al[2] compared the attendance rate at screening CT colonography with that at colonoscopy, correctly identifying via meta-analysis that the risk of *attendance* at screening was not significantly different between colonoscopy than CTC; but the directly opposite scenario, risk of *non-attendance* at screening, was significantly worse for colonoscopy than CTC. This highly counter-intuitive phenomenon, whereby simple reversal of the outcome of interest can lead to apparently different results, has been long-known[1,3,4], but is perhaps under-recognised. In this commentary, we aim to explain how this occurs, using examples in the literature where necessary, before making suggestions for how analysts and readers can mitigate the problem.

When considering a dichotomous outcome, researchers have to decide what represents the outcome of interest. Usually, this is obvious – death, myocardial infarction, or diagnosis of cancer are all unambiguous, clearly-defined, clinically-relevant endpoints, and are suitable outcomes for both component primary studies and subsequent meta-analysis. However, in some circumstances, it may be more difficult to define the relevant outcome. For example, when investigating screening, it is arguable whether attendance or non-attendance should be chosen as the outcome. Similarly, studies examining fertility treatment (for example) could choose to define successful conception as the outcome, or non-conception as an adverse outcome and thus the “risk event”. Why does this seemingly arbitrary choice matter? Because, for meta-analysis under certain circumstances, it can be extremely important depending on the analysis method chosen[3]. Analysing relative risks requires caution when the prevalence of the outcome varies across the component studies.

A summary statistic for meta-analysis is generated by pooling the individual estimates of the effects observed in the component primary studies. For binary outcomes, these are usually expressed as a relative risk (RR, also called the risk ratio) or an odds ratio (OR). Although exact methods for pooling component studies individual RRs or ORs vary, in essence meta-analysis assigns a “weight” to each component study based on how precisely the outcome measure can be estimated; and, when random effects meta-analysis is used, an adjustment for variation between studies[5]. The weighting given to an individual study determines its influence over the final meta-analysis pooled summary statistic; larger weightings exert greater effect. For fixed effect meta-analysis, individual study weights are affected primarily by the 95% confidence interval around the point estimate of the RR (or OR) – larger weightings are given to studies with narrower confidence intervals than the other studies in the meta-analysis.

This process seems logical – studies whose individual results are more precise should exert more effect on the final meta-analysis outcome measure. In random effects meta-analysis, smaller studies have larger relative weight, as the meta-analysis aims to estimate the average effect across all studies (rather than assuming there is an underlying “standard” effect that should be the same across all studies). Nonetheless, the width of the 95% confidence intervals around the risk estimate from each individual study still influences the final weight assigned even in random effects meta-analyses. Usually, this is not problematic; however, using relative risks can introduce unpredictable behaviour in meta-analysis when the prevalence of the risk event varies across the studies. Specifically, it can result in very different weights being assigned to studies that are otherwise similar. For example, consider the first forest plot presented by Zhu et al[2] (Figure 2a of their article; redrawn here for convenience).

We can see here that the two large studies by Stoop et al[6] (8844 patients) and Sali et al[7] (5861 patients) receive the largest weightings, of 23.9% and 22.3% respectively. However, You et al[8], which randomised only 131 patients, receives almost the same weighting, at 21.7%; and more than Scott et al[9] (weighting 17.4% for a sample of 709 patients) and the MACS group[10] (weighting 14.7% for a sample of 429 patients). This would not matter if individual study results were identical, but they are not (which, after all, is why we perform meta-analysis). For example, the RR for You et al is less than 1.0, but greater than 1.0 for Stoop et al. Accordingly, although the two largest studies show a clearly significant effect in favour of CTC, the overall meta-analytic point estimate is not significant at the 5% level (p=0.07) because it is "dragged down" by the weighting ascribed to smaller studies with conflicting findings. How has this apparently counterintuitive situation occurred, whereby a small study of 131 patients receives weighting virtually equivalent to a 6000 patient RCT?

Relative risks for randomised trials are a simple concept; the probability (or risk) of the outcome in one trial arm is compared to the probability of the same outcome in the other arm, expressed as a ratio. If 200 of every 1000 patients die with placebo (200/1000 = 0.2) and 150 die with treatment (150/1000 = 0.15), the relative risk of death is 0.15 / 0.20 = 0.75 (95%CI 0.62 to 0.91), strong evidence supporting treatment. The alternative is to calculate the odds ratio[11], which is, exactly as the name suggests, the ratio of the *odds* of the outcome in each study arm (rather than the probability). In this example, the odds of death with placebo are 200:800 i.e. 1:4 or 0.25, versus odds of death with treatment of 150:850 i.e. 0.176. The odds ratio is therefore 0.176 / 0.25 = 0.71 (95%CI 0.56 to 0.89). Little has changed to our conclusion – but the absolute value of the outcome metric and its confidence intervals are different.

Relative risks are often preferred to odds ratios because they are simpler to understand[12-14]. However, they have an unfortunate statistical property – their 95% confidence interval depends greatly on how frequently the outcome occurs. The commoner the outcome (i.e. the

greater its prevalence), the narrower the 95% confidence interval. Consider a trial in which the outcome (e.g. death) occurs in 10 of 100 untreated patients, and 9 of 100 treated patients; the RR is 0.90 and the 95% confidence interval is very wide, at 0.38 to 2.12. However, now consider a trial in which death is common, occurring in 50 of 100 untreated patients and 45 of 100 treated patients; the RR remains 0.90, but the 95%CI is far narrower, at 0.67 to 1.20. As the outcome becomes increasingly common, the 95%CI for the RR narrows progressively; if 90 of 100 untreated patients were to die versus 81 of 100 treated patients, the result is RR = 0.90, 95%CI 0.80 to 1.01. This explains the apparent discrepancy in our example – the prevalence of the outcome (attendance at screening) was far higher for You et al (80.3% for colonoscopy, 76.9% for CTC) than for all other studies (which ranged from 14.8% to 33.6%). This likely happened because You et al pre-selected their participants via an expression of interest in screening. The 95% confidence interval for the relative risk is narrower than expected and the study is weighted accordingly, despite randomising far fewer patients, fewer attendances at screening, and fewer non-attendances than Scott et al (i.e. all event categories were smaller).

Moreover, this effect can cause bizarre results when it is unclear in which “direction” we should define the outcome. For example, consider what happens if we choose to reverse the outcome categories, so that *non-attendance* at screening becomes the “risk event” (rather than attendance). This generates the forest plot shown in Figure 2.

Now we can see You et al has wide confidence intervals (0.61 to 2.26) and a low weighting in the meta-analysis (1.05% for RR of missed appointments compared to 21.07% for RR of attendance). Meta-analysis of RR for missed appointments is 0.92 with 95%CI 0.85 to 0.98, p = 0.01, conventionally significant at the 5% level. Therefore, different conclusions can be drawn from the same data, depending on whether attendance or non-attendance is defined as the “risk event”. Both analyses have been performed correctly, but their results vary

because 95%CIs for relative risks are not symmetric with respect to what is defined as the outcome[4]. This is clearly highly counterintuitive for clinical decision-making.

Although we are not aware of this phenomenon being described for previous radiological meta-analyses, it has been reported in other scenarios. For example, describing a meta-analysis comparing eradication of *Helicobacter pylori* for non-ulcer dyspepsia versus placebo for trials conducted before 2000[4], Deeks observed the same phenomenon. When considering the outcome to be "ongoing dyspepsia", the RR for treatment was 0.92 (95%CI 0.85 to 0.99), significant at the 5% level. However, if the outcome definition were to be reversed (i.e. the outcome is "no dyspepsia"), the RR for treatment was 1.28 (95%CI 0.92 to 1.77), no longer significant at the 5% level (as more trials have been published, the benefit of H. pylori eradication has become clear). It is clear that, for meta-analyses using relative risks, there are in fact two possible summary risks – one for benefit (sometimes called $RR_B$), and one for harm, $RR_H$[4]; and there is no easily-predictable mathematical relationship between the two.

So, how can the problem be resolved? Firstly, we can avoid over-reliance on the arbitrary 5% threshold to define significance, and inspect the data itself, which should prevent spurious conclusions. A second option is to use odds ratios, the 95% confidence intervals for which are unaffected by prevalence, or by which event category is defined as the outcome. For example, Figure 3 shows the same data as in figures 1 and 2, but now presented as odds ratios. Irrespective of whether we define attendance or non-attendance as the "risk event", study weights are assigned consistently and each summary estimate is simply the reciprocal of the other, which seems intuitive if we have reversed the outcome.

It is important to note that, under most circumstances, it is clear in which direction the "risk event" should be specified (i.e. what constitutes the outcome of interest); and empirical data

shows that defining an adverse event as the outcome typically gives more consistent results and is preferred, certainly for preventative (rather than therapeutic) interventions[4]. Generally speaking, whichever is the less common state is usually preferable as being the "risk event". Moreover, relative risks are much easier to interpret than odds ratios[12-14], meaning they are often preferred for this reason alone. Fortunately, the situation that we have outlined above is rare, since it depends on both (a) a wide range of prevalences of the outcome of interest occurring in the component studies and (b) effect sizes varying between these different studies, which may not always be the case. Nonetheless, using relative risks may introduce a bias towards larger weights being assigned to smaller, early-phase RCTs (which typically are targeted to high prevalence scenarios in order to maximise event rates) than to larger, pragmatic RCTs (which aim to recruit more representative patient populations, sometimes including many without the outcome of interest). It seems fundamentally wrong that the outcome prevalence can influence weighting within meta-analysis, so that smaller studies, paradoxically recruiting higher-risk, less-representative patients, can outweigh larger studies. For example, imagine six studies comparing death rates in a series of placebo-controlled trials, all of which have the same average effect size when expressed as a relative risk (RR = 0.9), but recruiting different numbers of participants and with two differing death rates, 26% and 70%. Table 1 summarises these hypothetical data:

Table 1: Hypothetical data for six randomised trials of varying sizes and with varying event rates.

| Study name | Treatment | | Placebo | | Total number of participants | Average event rate (death rate) |
|---|---|---|---|---|---|---|
| | Died | Survived | Died | Survived | | |
| A | 600 | 300 | 600 | 210 | 1710 | 70% |
| B | 600 | 1800 | 600 | 1560 | 4560 | 26% |
| C | 300 | 150 | 300 | 105 | 855 | 70% |
| D | 300 | 900 | 300 | 780 | 2280 | 26% |
| E | 120 | 60 | 120 | 42 | 342 | 70% |

| | | | | | | |
|---|---|---|---|---|---|---|
| F | 120 | 360 | 120 | 312 | 912 | 26% |

Intuitively, we might expect studies B and D to contribute the greatest weight to a meta-analysis, since they are the largest; however, meta-analysis using relative risks, whether with fixed or random effects, shows this is not the case (figure 4).

Despite identical relative risks for all studies, the largest have wider 95% confidence intervals and thus contribute smaller weights to meta-analysis, because the prevalence of the outcome (death) is lower than in the smaller studies. The largest study (Study B) has less weight within meta-analysis than the much smaller Study C, despite its raw data contributing more patients in all categories (i.e. died or survived, for both treatment and placebo). Therefore, as outlined by the Cochrane handbook[15], it "may be wise to plan to undertake a sensitivity analysis to investigate whether choice of summary statistic (and selection of the event category) is critical to the conclusions of the meta-analysis" where component study prevalence is variable; and consider using odds ratios rather than relative risks for the primary analysis (bearing in mind the difficulties with their subsequent interpretation). A further option, adopted by Zhu et al[2], is to report both relative risks i.e. for both definitions of the outcome. This is probably only appropriate when it is arguable in which direction the outcome should be specified (e.g. neither is clearly a negative or harmful event, or much less common – in which case, the rarer and/or negative event should be specified as the outcome).

In summary, we urge readers of meta-analyses, and researchers themselves, to consider carefully their choice of summary statistical measure when analysing dichotomous outcomes. Although relative risks are commonly chosen for simplicity, where outcome prevalence varies greatly between component studies, and particularly where it is not clear

which category of outcome should be regarded as the event, researchers should exercise caution and follow the Cochrane handbook guidance outlined above.

**Figure Legends:**

Figure 1: Forest plot similar to that generated by Zhu et al via random effects meta-analysis using relative risks for attendance at screening; larger values imply greater attendance at CTC when compared to colonoscopy. The summary estimate is 1.26 (95%CI 0.98 to 1.63), p = 0.07, not statistically significant at a 5% level (despite the two largest trials finding a significant result in favour of CTC).

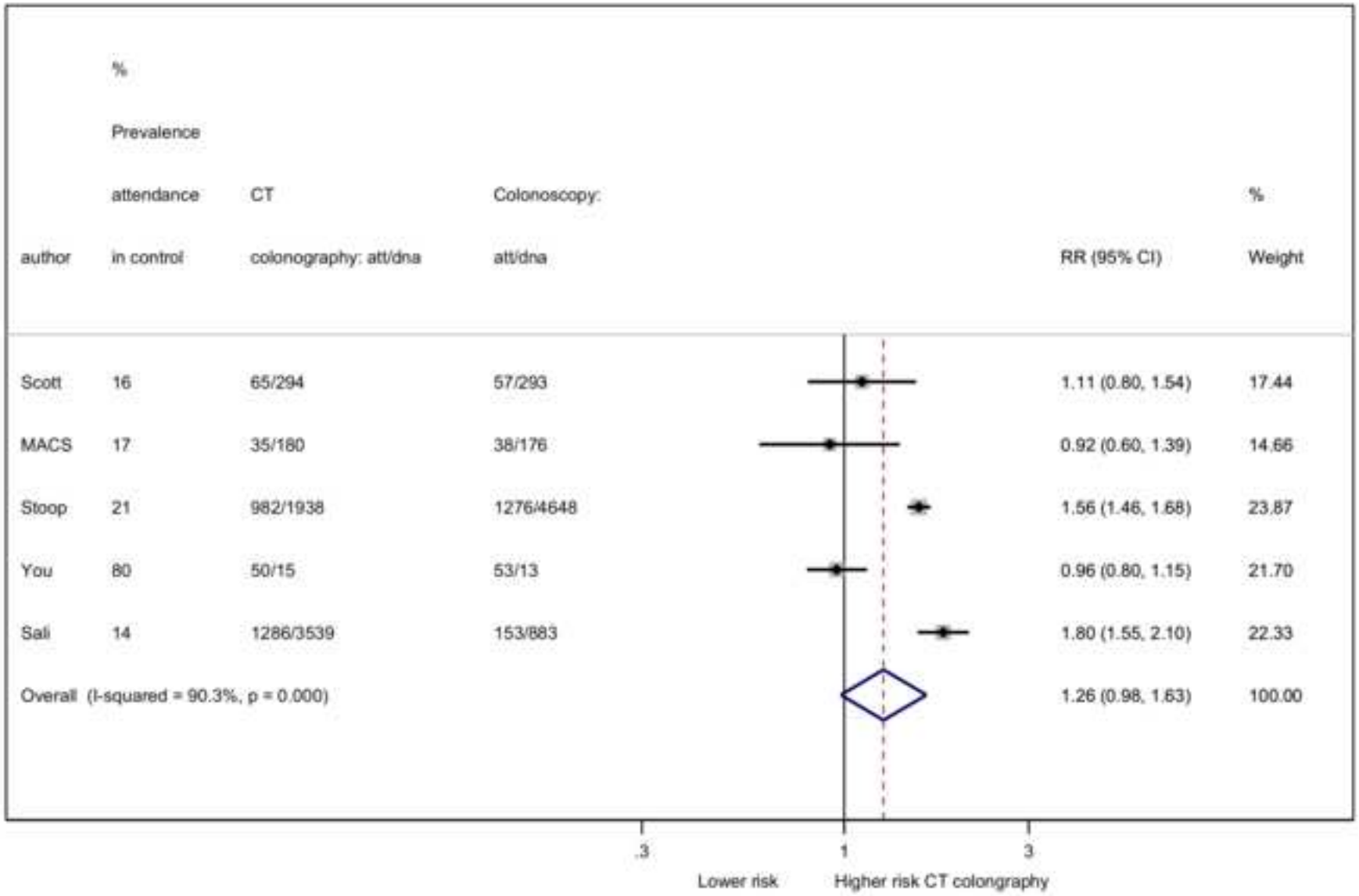
Figure 2: Forest plot generated using identical data for that in Figure 1, again via random effects meta-analysis using relative risks, but reversing the category of the outcome, such that relative risks indicate the risk of non-attendance at screening CTC vs colonoscopy. The summary estimate is 0.92 (95%CI 0.85 to 0.98), p = 0.01.

Figure 3: Forest plots generated using identical data for that in Figures 1 and 2, again via random effects meta-analysis but now using odds ratios. Figure 3a shows attendance at screening being the “risk event” (or outcome), whereas figure 3b uses non-attendance as the outcome. The study weights are the same in each analysis, the forest plots are simple “mirror images” and the summary odds ratios are reciprocals of each other.

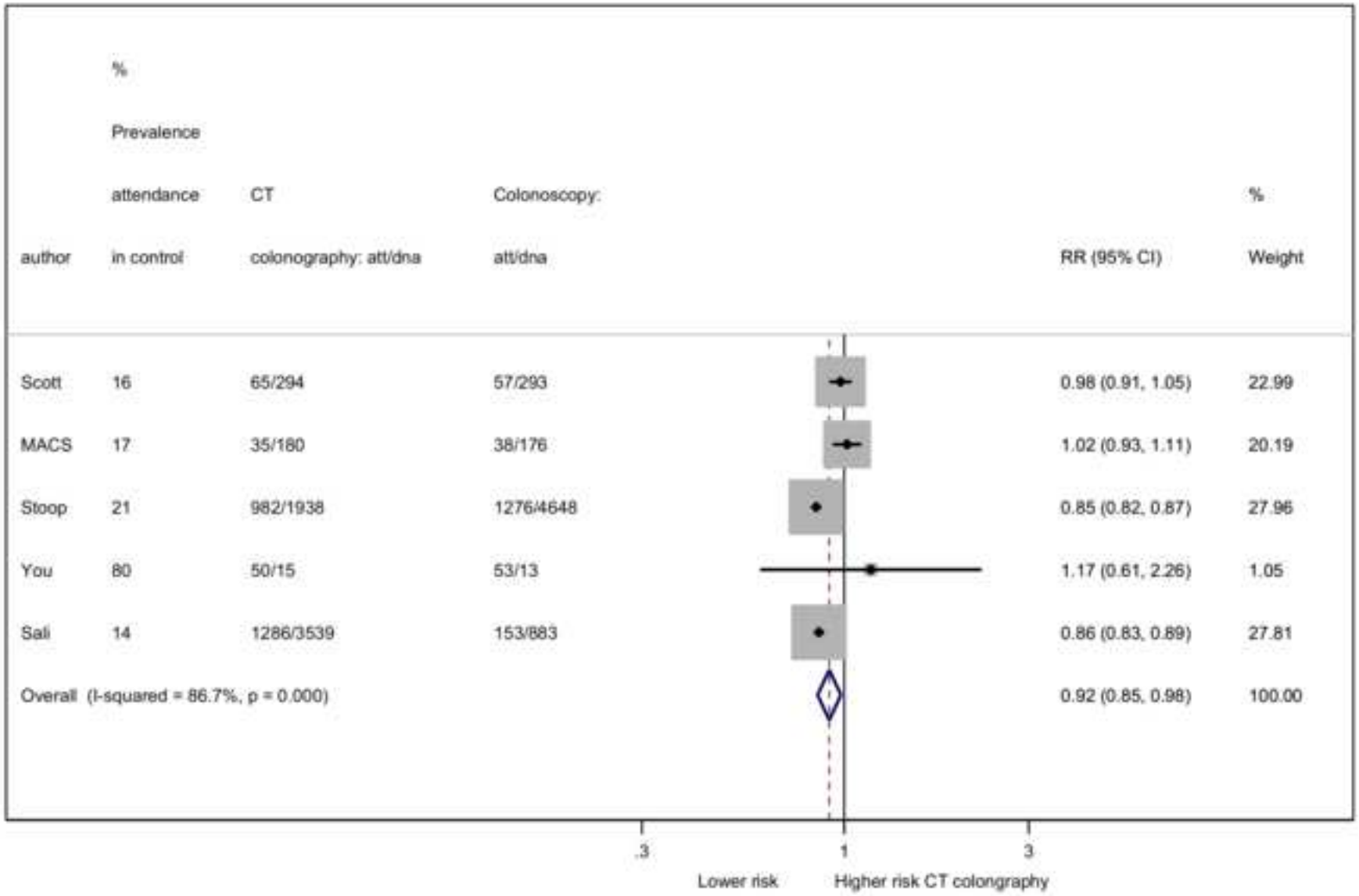
Figure 4: Forest plot generated using the hypothetical data from Table 1 and for random effects meta-analysis using relative risks. The same would be seen for a fixed effect meta-analysis (since these hypothetical studies all have the same effect size, there is no between-study variance and so the weights are the same for random and fixed effects meta-analyses).

## References

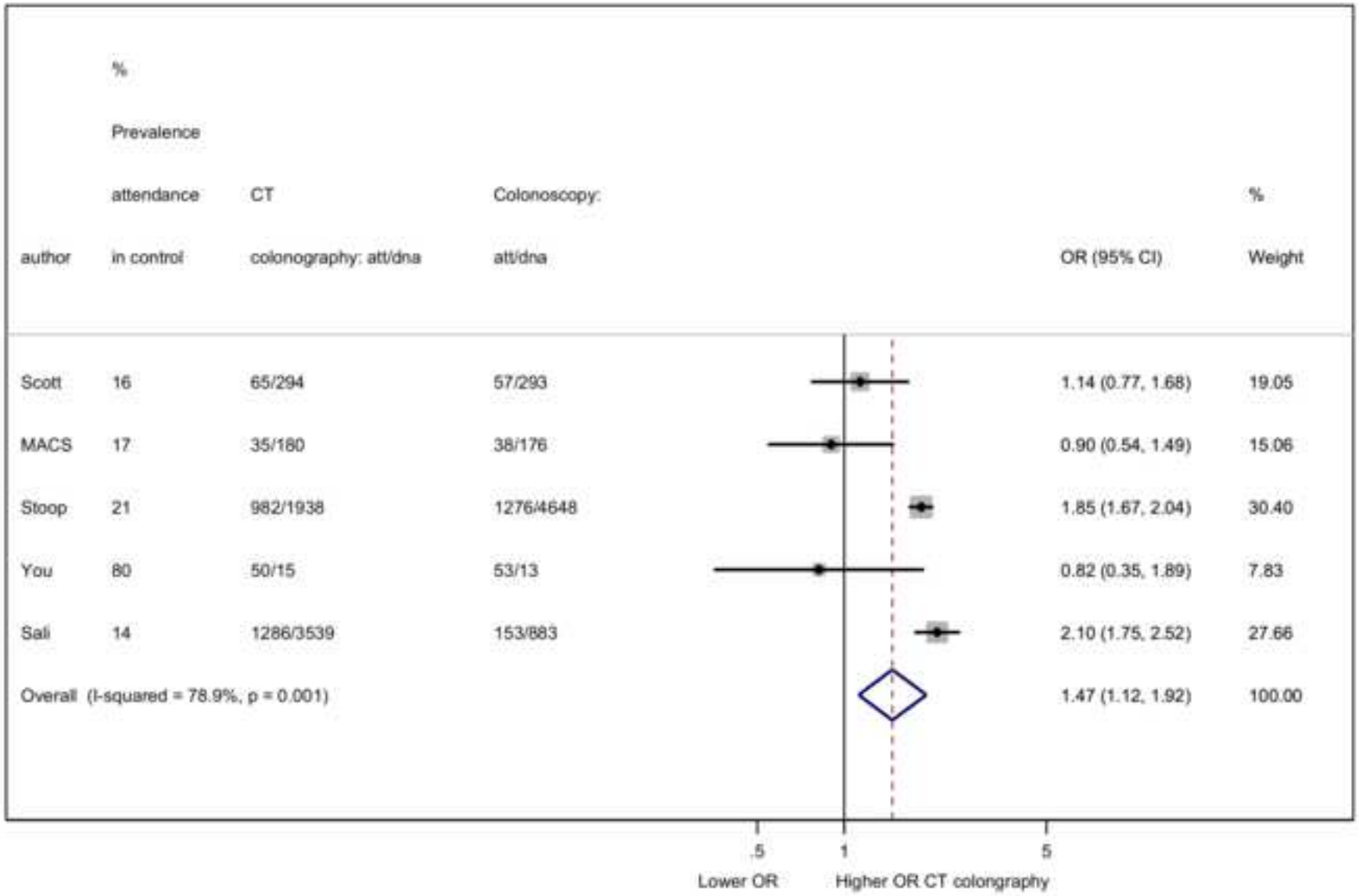

1. Cox DR, Snell EJ. Analysis of binary data. 2nd ed. ed: Chapman and Hall; 1989.
2. Zhu H, Li F, Tao K, et al. Comparison of the participation rate between CT colonography and colonoscopy in screening populations: A systematic review and meta-analysis of randomized controlled trials. *British Journal of Radiology* 2019; **92**: 20190240.
3. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med* 2000; **19**(13): 1707-28.
4. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002; **21**(11): 1575-600.
5. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010; **1**(2): 97-111.
6. Stoop EM, de Haan MC, de Wijkerslooth TR, et al. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. *Lancet Oncology* 2012; **13**(1): 55-64.
7. Sali L, Mascalchi M, Falchini M, et al. Reduced and Full-Preparation CT Colonography, Fecal Immunochemical Test, and Colonoscopy for Population Screening of Colorectal Cancer: A Randomized Trial. *JNCI J Natl Cancer Inst* 2016; **108 (2) (no pagination)**(djv319).
8. You JJ, Liu Y, Kirby J, Vora P, Moayyedi P. Virtual colonoscopy, optical colonoscopy, or fecal occult blood testing for colorectal cancer screening: results of a pilot randomized controlled trial. *Trials [Electronic Resource]* 2015; **16**: 296.
9. Scott RG, Edwards JT, Fritschi L, Foster NM, Mendelson RM, Forbes GM. Community-based screening by colonoscopy or computed tomographic colonography in asymptomatic average-risk subjects. *American Journal of Gastroenterology* 2004; **99**(6): 1145-51.
10. Multicentre Australian Colorectal-neoplasia Screening G. A comparison of colorectal neoplasia screening tests: a multicentre community-based study of the impact of consumer choice. *Med J Aust* 2006; **184**(11): 546-50.
11. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ* 2000; **320**(7247): 1468.
12. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead? *BMJ* 1998; **316**(7136): 989-91.
13. Cummings P. The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med* 2009; **163**(5): 438-45.
14. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evidence Based Medicine* 1996; **1**(6): 164-6.
15. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. 2011.



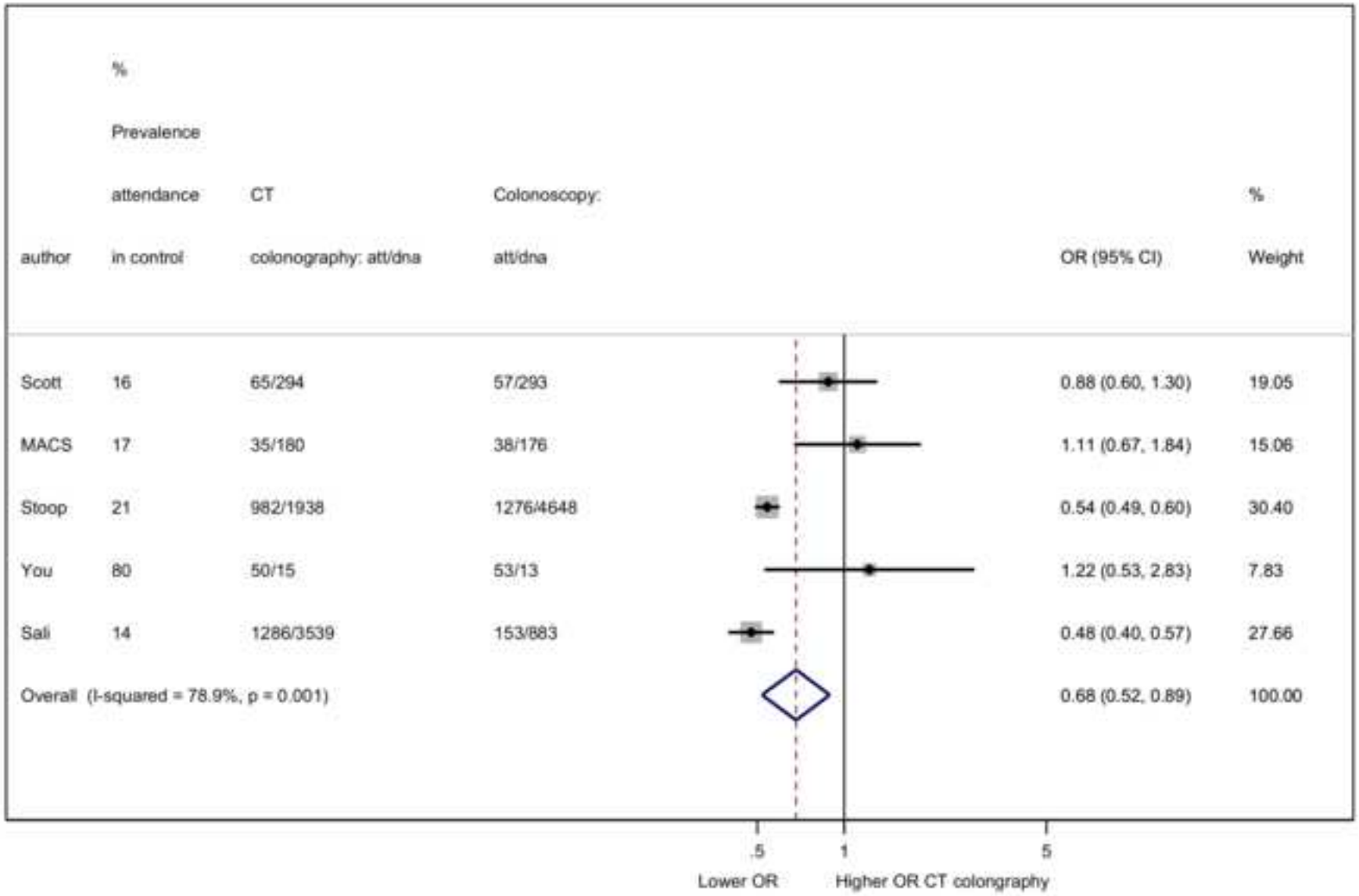

Attendence: risk ratio




Missed appointments: risk ratio




Attendence: odds ratio




Missed appointments:odds ratio