Synthesis of data from trials of interventions designed to change health behaviour; a case study

Sarah Ann Rhodes[1], Sofia Dias[2], Jack Wilkinson[1], and Sarah Cotterill[1]

[1]The University of Manchester
[2]University of York

September 25, 2021

Abstract

Many complex healthcare interventions aim to change the behaviour of patients or health professionals, e.g. stopping smoking or prescribing fewer antibiotics. This prompts the question of which behaviour change interventions are most effective. Synthesising evidence on the effectiveness of a particular type of behaviour change intervention can be challenging because of the high levels of heterogeneity in trial design. Here we use data from a published systematic review as a case study and compare alternative methods to address this heterogeneity. One important sources of heterogeneity is that compliance to a desired behaviour can be measured and reported in a variety of different ways. In addition, interventions designed to target behaviour can be implemented at either an individual or group level leading to trials with varying layers of clustering. To handle heterogeneous outcomes we can either convert all effect estimates to a common scale (e.g. using standardised mean differences) or have separate meta-analyses for different types of outcome measure (binary and continuous measures).To address the clustering structure, adjusted standard errors can be used with the inverse variance method, or weights can be assigned based on a consistent level of clustering, such as the number of healthcare professionals. A graphical method, the albatross plot utilises reported p-values only, and can synthesise data with both heterogeneous outcomes and clustering with minimal assumption and data manipulation. Based on these methods, we reanalysed our data in four different ways and have discussed the strengths and weaknesses of each approach.

Synthesis of data from trials of interventions designed to change health behaviour; a case study.

Sarah Rhodes[1]

Email: sarah.a.rhodes@manchester.ac.uk

Sofia Dias[2]

Email: sofia.dias@york.ac.uk

Jack Wilkinson[1]

Email: jack.wilkinson@manchester.ac.uk

Sarah Cotterill[1]

Email: sarah.cotterill@manchester.ac.uk

[1]Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

[2]Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK

Abstract

Many complex healthcare interventions aim to change the behaviour of patients or health professionals, e.g. stopping smoking or prescribing fewer antibiotics. This prompts the question of which behaviour change interventions are most effective. Synthesising evidence on the effectiveness of a particular type of behaviour change intervention can be challenging because of the high levels of heterogeneity in trial design. Here we use data from a published systematic review as a case study and compare alternative methods to address this heterogeneity. One important sources of heterogeneity is that compliance to a desired behaviour can be measured and reported in a variety of different ways. In addition, interventions designed to target behaviour can be implemented at either an individual or group level leading to trials with varying layers of clustering.

To handle heterogeneous outcomes we can either convert all effect estimates to a common scale (e.g. using standardised mean differences) or have separate meta-analyses for different types of outcome measure (binary and continuous measures).To address the clustering structure, adjusted standard errors can be used with the inverse variance method, or weights can be assigned based on a consistent level of clustering, such as the number of healthcare professionals. A graphical method, the albatross plot utilises reported p-values only, and can synthesise data with both heterogeneous outcomes and clustering with minimal assumption and data manipulation.

Based on these methods, we reanalysed our data in four different ways and have discussed the strengths and weaknesses of each approach.

Keywords

Behaviour change, evidence synthesis, trials.

Background

In healthcare, many complex interventions are designed with the aim of changing the behaviour of individuals or groups of individuals. When designing new interventions, it is helpful to know which behaviour change techniques are most effective, and in which context. The behaviour change technique taxonomy[1] has identified and classified 93 different behaviour techniques, and each of these may be used alongside other techniques to form a complex intervention. The types of behaviours that these interventions could be targeting are numerous and varied; health behaviours such as eating a low calorie diet, or ceasing smoking; or clinical behaviours such as following government guidelines, prescribing drugs or washing hands. When summarising behaviour change research, one option would be to consider the effect of a specific intervention on a specific behaviour, but the large number of targeted behaviours would lead to a huge number of potential systematic reviews (or comparisons within a systematic review); each aiming to answer a different question, but unlikely to have enough statistical power to do so. In addition, it may be hard to interpret evidence from multiple systematic reviews of similar interventions that report conflicting conclusions, and present evidence in different ways. As described by Melandez-Torez[2], there are situations where it makes sense to group together interventions as 'clinically meaningful units' with a similar expected 'theory of change'. In terms of behaviour change techniques, it can be informative to combine evidence to answer a broad question about how well a particular behaviour change technique (or group of techniques) has performed, on average, on any type of behaviour, and to use this information to identify which techniques are effective. This can be supplemented with analysis of effect moderators, to identify the contexts in which the technique is more or less effective.

Interventions to change healthcare professional (HCP) behaviour can be designed to target the individual HCP, or team of HCPs. Trials of this type of intervention can vary in terms of the unit of randomisation which can be either the individual HCP, or a group of HCPs such as those working within the same site (surgery, nursing home, ward, hospital). The unit of analysis in these trials can also vary and is not necessarily the same as the unit of randomisation; for example with randomisation at the level of GP surgery but with data recorded for each individual patient. The outcomes could be measured using a variety of denominators, such as the individual patient (e.g. binary measure of whether a test was ordered), individual HCPs (e.g. number

of tests ordered per GP), or at site-level (e.g. proportion of patients with an appropriate test order on a hospital ward). These multiple and varied layers need to be considered in terms of adjustment for clustering, combination of data and interpretation of results.

There are several proposed methods of summarising mixed measures of behavioural outcome. Higgins et al.[3] provide an overview of methods to synthesise quantitative evidence in systematic reviews of complex health interventions. They describe and compare a number of graphical methods to combine different outcomes; as well as synthesis methods using effect size estimates, which are suitable for complex interventions. One approach is to combine effect sizes (standardised mean differences (SMDs)) using standard errors to derive weights for the studies in meta-analysis. In addition to allowing for different measurements (both binary and continuous) to be combined, this approach can accommodate a mixture of individually randomised and cluster randomised trials using weights based on adjusted standard errors. In some systematic reviews, binary measures of the same outcome are analysed and reported separately from continuous ones[4,5]. An alternative approach[6] to using weights based on standard errors is to use study weights based on the number of health professionals included in the study. The albatross plot[7] is a graphical method which allows synthesis of summary data in a variety of formats, using only p-values plotted against sample size; this can be used to assess the consistency of results visually and allows estimation of average effect sizes.

Aims

Our aim was to examine methods for conducting meta-analysis in the context of heterogeneous behavioural outcome measures with clustering using a case study. We have applied different methods to data from the SOCIAL systematic review[8-10] to illustrate the methods and examine their strengths and weaknesses.

Dataset

SOCIAL [89,10]was a systematic review of randomised controlled trials, looking at the effect of social norms interventions on the clinical behaviour of health care workers, where a social norms intervention is defined as 'an intervention which aims to change the behaviour of an individual by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person'(Tang, 2021, p.2) This review looked at the effects of any social norms intervention on any type of clinical behaviour, and aimed to answer an overall question about the effectiveness of social norms interventions, as well as more specific questions related to different types of intervention, settings, contexts and behaviour.

The SOCIAL systematic review included 102 unique trials that assessed the effect of a social norms intervention on the clinical behaviour of health workers. For ease of presentation, here we focus on a subset: 16 trials that assessed the effect of 'credible source' interventions either alone or alongside other interventions. A credible source intervention provides communication either in favour of or against a particular behaviour by a person generally agreed on as credible with the aim of persuading the recipient[1]. For example Hallsworth [11] include a persuasive letter from the Chief Medical Officer in their intervention to reduce antibiotic prescriptions amongst high prescribing GPs.

Note that 2 of these 16 trials had more than two arms that tested the effect of a credible source intervention, so there were 18 different comparisons included. Table 1 shows the units of randomisation and analysis and how they vary by study.

The SOCIAL review found that social norms interventions appeared to be an effective method of changing the clinical behaviour of healthcare workers, with credible source interventions appearing to be most effective on average[10].

Analysis

All analyses were performed using STATA 14.0[12]. We have reported results of both fixed and random effects meta-analysis.

Where some summary data were missing for the reported trial results, we have imputed missing information (e.g. standard deviations or intra-cluster correlation coefficients (ICCs)) using other information in the trial

paper or values from similar trials[13]. Sensitivity to imputed values was assessed by imputing a range of different values.

Where the reported outcome data were from either an individually randomised trial or a cluster trial where the results had already been adjusted for clustering by the unit of randomisation, the standard errors were utilised without adjustment. Where adjustment for clustering was required the standard error was multiplied by the square root of the design effect (DE); this requires the average cluster size (M) and the intra-class correlation coefficient (ICC)[13].

$DE = 1+(\text{M-1})\text{ICC}$

Where possible we report the I^2 statistic as a measure of heterogeneity [13]. The $\mathbf{I^2}$ statistic estimates the percentage of variation across studies that is due to study heterogeneity rather than chance.

Method 1: Standardised Mean Differences, weights based on adjusted standard error

This method is commonly used, including in the SOCIAL systematic review[8] and other reviews[14-17]. This method is simple to use, it utilises information that is generally reported, and it can be performed using standard statistical software. All reported measures of intervention effect are converted into an approximation of the standardised mean difference (SMD) using the formulae in Table 2[18,19].

The formula for a standardised mean difference for a continuous outcome (Table 2) refers to Cohen's d for ease of calculation. As an alternative Hedges g may be used [19]which allows a correction for small sample size.

We applied these methods to the SOCIAL meta-analysis using the inverse of the squared adjusted standard error as weights (inverse variance method [13]).

Where a trial reported both a continuous and binary outcome measure with appropriately adjusted standard errors, we utilised the continuous measure but also calculated the SMDs and standard errors using the binary measure to check for anomalies. Note that rules such as this should be pre-specified to avoid post-hoc decisions that could introduce bias.

Method 2: Separate analyses for binary and continuous outcomes, weights based on adjusted standard errors

In this method two separate analyses are produced for the same outcome; one for those that were reported as binary measurements and one for those that were reported as continuous measurements. This method has been used in a number of systematic reviews of health behaviour change[4,5,16]. This method requires very little data manipulation and adopts a conservative approach to heterogeneity by keeping the two types of outcome separate.

For illustration we performed meta-analysis on the SOCIAL data using odds ratios for binary data and standardised mean differences for continuous data, using the inverse variance method with weights based on adjusted standard errors.

Outcomes were reported as continuous measures on a variety of different scales; therefore they were converted to standardised mean differences as above [20] . If all continuous measures had been measured on the same scale, e.g. mean percentage on a scale of 0 to 100, they could be meta-analysed using means and standard deviations. For meta-analysis of binary data, all summaries need to be converted to the same format (odds ratio, risk ratio or risk difference); it is recommended that this be chosen in advance at the protocol stage to avoid selective reporting. We chose odds ratios [18] here as they have certain desirable mathematical properties; their symmetrical nature would mean that an analysis where the outcome measure is 'compliance' or an analysis where the outcome measure is 'non-compliance' would lead to identical conclusions.

Sometimes trials report the same outcome measure in both binary and continuous formats – for example in a trial where the desired behaviour is 'test ordering'; summary data could be reported both in terms of the overall proportion of patients who had a test ordered, and the mean proportion of tests ordered by health

care professional. Where a trial has reported an outcome in both binary and continuous formats, we included both measures in the two separate meta-analyses. Note that when using this method, continuous and binary results may not be later combined together as this would lead to double counting of the same participants.

Method 3: SMDs, weighting by number of health care professionals

Where the population of interest is the health care professional, it may be desirable to weight the results by the number of health care professionals included[6,21]to aid population inference. This method utilises commonly reported summary information without adjustment for clustering. In this method of meta-analysis the studies are weighted by the number of health care professionals as an alternative to the commonly used inverse variance method which uses weights based on standard errors. As pointed out in table 4 there are weaknesses to this approach which we discuss below.

Not every trial in the SOCIAL review reported the number of health care professionals – for example a cluster trial where an intervention was directed at all staff on a hospital ward. Where no information was given about the number of health care professionals, Ivers et al.[6] used the number of practices/hospitals/communities instead, and we followed that method here. An alternative might be to estimate the number of health care professionals using data from similar studies – e.g. using mean number of GPs per surgery or mean number of nursing staff on a hospital ward. Note that this method needs to be combined with method 1 or method 2 above or an alternative way of summarising mixed outcome measures; here we combined it with method 1 to summarise standardised mean differences.

Method 4: Albatross plots

The albatross plot was first described by Harrison et al[7] and is also discussed in Higgins [22]. This method requires minimal data extraction or manipulation and allows data to be synthesised even in circumstances when outcomes are reported in multiple different formats or where no summary statistics are reported. Reported results are split into two groups according to the direction of effect; and then p-values are plotted against sample size. Where necessary, 1-sided p-values need to be converted to 2-sided p-values (or vice versa) to ensure consistency. An albatross plot allows us to combine outcome data that was reported in a variety of different ways, including from studies where only a p-value was provided. Under an assumption of normality, you would expect results corresponding to the same effect size to lie along a contour, with p-values generally getting smaller as sample size increases. Contours can be added to the plot for a range of different effect sizes based on standardised mean differences, mean differences, odds ratios or other summary of choice. Effect sizes can be estimated according to where the majority of points lie. We have added contours to represent standardised mean differences of 0.3, 0.6 and 0.9. Heterogeneity can also be explored visually by looking at how closely trials tend to group together along a particular contour.

Note that where p-values are obtained from studies that are clustered in some way, adjustment of sample size is necessary. One method of doing this is to calculate the effective sample size (E) using the sample size (S), the reported intra-class correlation coefficient (ICC) and the average cluster size (M) using the formula[23]

$$E = \frac{S}{1 + ICC \ \times (M - 1)}$$

An alternative is to replace the sample size with the number of health care professionals (or sites) as in method 3.

For illustration we produced a contour plot using the number of health care professionals (or sites) as the sample size (Figure 1)

Results

Method 1 produced pooled SMDs of 0.14 (95% CI 0.10 to 0.17) and 0.31 (95% CI 0.14 to 0.51) for the fixed and random effects results. There is a marked difference between the results for fixed and random effects;

with the fixed result having a smaller effect size and tighter confidence interval; this is because the fixed effects analysis gives more weight to large trials, which tended to have more modest effect sizes (Table 3).

Method 2 resulted in pooled OR 1.13 (95% CI 1.06 to 1.20) and pooled SMD 0.50 (95% CI 0.42 to 0.59) for fixed effects; OR 1.13 (95% CI 1.06 to 1.20) and SMD 0.92 (95% CI 0.11 to 1.73) for random effects. One study contributed data to both the odds ratio and SMD estimate. The method using odds ratios produced a far less heterogeneous result than that for the SMDs in this case but as they are from different sets of trials it is difficult to infer why.

Method 3 resulted in an SMD 0.57 (95% CI 0.50 to 0.64). This weighted average produced the narrowest confidence intervals for SMDs.

For Method 4, we can see from Figure 1 that all studies reported a positive effect so it is clear that, on average, credible source interventions seem effective. The fact that the points are not clustered around one particular contour line tells us that there is a high level of heterogeneity. Both large and small studies appear to be associated with very small p-values and large effect sizes, so there is little evidence of publication bias.

Three of the methods produced an SMD, which ranged from 0.14 to 0.57. All were statistically significant, suggesting that we can be reasonably confident that a positive effect exists, but less confident in estimating the size of the effect as it is sensitive to the method chosen.

Challenges

In Table 4 we summarise the strengths and weaknesses of the different approaches. In systematic reviews of complex interventions there is likely to be a large amount of heterogeneity due to differences in setting, population, intervention and study design. When combining different types of outcomes, measured and reported in a variety of different ways, heterogeneity due to outcome measurement also has to be a serious additional consideration. Our estimate of heterogeneity, I^2 for the SMD analyses ranged from 95.3% to 98.5% suggesting substantial heterogeneity. Exploration of heterogeneity is not the focus of the paper and has been discussed by a number of authors[24,25,26]. Sources of heterogeneity can be explored using methods such as subgroup analysis[27] and meta-regression[28] although these common approaches are subject to ecological fallacy, and superior approaches exist where sufficient data are available[29] . In contrast to a meta-analysis of a well-defined pharmaceutical intervention, where heterogeneity is generally seen as a nuisance, identifying the sources of the heterogeneity is often a key research questions when synthesising data from complex interventions.

Some authors have expressed concerns about the use of SMDs in meta-analysis. The SMD estimates the average improvement in outcome per SD on whatever scale that outcome is measured on; as Greenland[30] points out the SD measured within a trial is likely to be different to the population SD and will vary according to the design features of the trial (e.g. inclusion/exclusion criteria). Trials are often designed to minimise variability and therefore SDs reported are likely to be smaller than the SD in the target population, leading to an overestimate of the treatment effect of interest. Another problem, as discussed by Senn[31], that is especially pertinent here, is that the SD will depend on the measurement error, and since we have lots of different measurement scales we will have lots of different measurement errors; this means that you could get lots of different SMDs even if the treatment effect was the same in each study.

In an attempt to combine all available information we have converted odds ratios into SMDs using the methods described by Chinn[32]. This method provides an estimate of the SMD from an odds ratio using the assumption that the odds ratio has come from a dichotomy of a normally distributed continuous variable; this may be a poor estimate when this assumption is not true. Sanchez-Meca[33] compares alternative indices to combine continuous measures with dichotomies and show that this method slightly underestimates the SMD. Our conclusions were unchanged when binary and continuous data were analysed separately, but the SMDs estimated from continuous data alone were considerably higher than those when binary data were combined so it is possible that by converting odds ratios to SMDs we were underestimating the true treatment effect in this context.

Varying units of randomisation and analysis lead to difficulties both in terms of synthesis methods and interpretation. One of our reported methods (Method 3) aims to apply consistent weighting based on the number of health care professionals to allow inference about a consistent population; however this leads to other problems. Weights based on sample size do not take into account the variability of the data, essentially assuming a constant standard deviation across all trials. In their simulation study, Marin-Martinez and Sanchez-Meca[34] show that weighting by the inverse variance yields less biased results than weighting by sample size. Complexity is added when a review wishes to combine evidence from different types of trial design[35,36]. Individually randomised trials, cluster randomised trials and stepped wedge trials are all useful in answering questions about behaviour change interventions targeted at health care professionals, but you would not necessarily expect the SMD (effect size) to be consistent across each type of trial due to the different units of analysis (and therefore different underlying SDs)[37,38]. Some consensus among trialists of health professional behaviour change interventions, in the form of a core outcome set[39]would be useful for future systematic reviews. Consistency in terms of outcomes used, unit of analysis and format of outcome reporting is desirable. In addition, we may want to separate out the effect on the health care provider from the effect on the individual patient; this would require individual participant data and multilevel modelling[24].

Some trials used in this analysis have reported 'mean percentage compliance' or similar – e.g. the percentage of occasions a test was ordered, averaged over a group of GPs. This measurement is bounded between 0% and 100% and therefore cannot be considered truly continuous. Inference methods (meta-analysis of SMDs) used here assume continuity and normality and are likely to perform poorly where results are close to the boundaries (0% and 100%). We performed additional sensitivity analyses removing trials where the mean compliance was between 0% and 20% or between 80% and 100%; and results appeared robust. Alternative methods to analyse proportions include those suggested by Miller [40]and Stijnen et al.[41]and these may be preferable when meta-analysing proportions alone.

We acknowledge all of these challenges and feel that conclusions based on any of the methods presented here need to be very cautious. However we feel that there are occasions where the combination of mixed outcomes is still warranted, but should be accompanied with appropriate sensitivity analyses and caveats.

Conclusions

Systematic reviews of complex behaviour change interventions in healthcare may include a heterogeneous set of studies in terms of content, context, trial design and setting. The measures of behaviour change may also vary which leads to difficulty in attempts to synthesise the data, as well as increased heterogeneity.

In this paper we have presented 4 different methods for combining behavioural outcome measures from trials, described the strengths and weaknesses of each method, and the problems inherent with combining heterogeneous outcome measures with mixed levels of clustering. Each of the methods presented has advantages and disadvantages, summarised in table 4, and we recommend that reviewers chose their methods carefully based on the needs of their review, and plan methods and data conversion policies in advance to avoid selective reporting. We observed that for our data, conclusions would remain robust regardless of the methods of analysis chosen; however the estimated magnitude of the treatment effect varied quite markedly according to the method chosen. We view the methods presented as useful when trying to convert all outcome measures to the same scale and to provide an overall summary, but results should be interpreted extremely cautiously given the limitations. We would recommend that results are used as an aid in summarising the evidence and generating future hypotheses rather than to infer future effects.

Funding

This project is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research, reference 17/06/06 - The impact of social norms interventions on clinical behaviour change among healthcare workers: a systematic review. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article

Ethics approval (include appropriate approvals or waivers)

This article utilises only summary data from published trials. It is exempt from the need for ethical approval.

Availability of data and materials (data transparency)

Data to be stored on Figshare (DOI will be made available)

Code availability (software application or custom code)

Code to be stored on Figshare (DOI will be made available)

Authors' contributions

SR conceived the idea for the article, performed analyses and wrote the draft manuscript. SC led the original systematic review, developed the idea for the article and wrote sections of the article. SD and JW provided methodological input and substantially edited the article.

References

1. Michie S, Richardson M, Johnston M, et al. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine* . 2013;46(1):81-95. doi:10.1007/s12160-013-9486-6

2. Melendez-Torres GJ, Bonell C, Thomas J. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Med Res Methodol* . 2015;15:47-47. doi:10.1186/s12874-015-0040-z

3. Higgins JPT, López-López JA, Becker BJ, et al. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health* . 2019;4(Suppl 1):e000858. doi:10.1136/bmjgh-2018-000858

4. Davey P, Marwick CA, Scott CL, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database of Systematic Reviews* . 2017;(2)doi:10.1002/14651858.CD003543.pub4

5. Vaona A, Banzi R, Kwag KH, et al. E-learning for health professionals. *Cochrane Database of Systematic Reviews* . 2018;(1)doi:10.1002/14651858.CD011736.pub2

6. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews* . 2012;(6)doi:10.1002/14651858.CD000259.pub3

7. Harrison S, Jones HE, Martin RM, Lewis SJ, Higgins JPT. The albatross plot: A novel graphical tool for presenting results of diversely reported studies in a systematic review.*Research Synthesis Methods* . 2017;8(3):281-289. doi:10.1002/jrsm.1239

8. Cotterill S, Powell R, Rhodes S, et al. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. *Systematic Reviews* . 2019/07/18 2019;8(1):176. doi:10.1186/s13643-019-1077-6

9. Cotterill S, Tang MY, Powell R, et al. Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis. 2020;8:41. doi:10.3310/hsdr08410

10. Tang MY, Rhodes S, Powell R, et al. How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis. *Implementation Science* . 2021/01/07 2021;16(1):8. doi:10.1186/s13012-020-01072-1

11. Hallsworth M, Chadborn T, Sallis A, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. *Lancet* . Apr 23 2016;387(10029):1743-52. doi:10.1016/S0140-6736(16)00215-4

12. Stata C. *Stata release 14* . 2015.

13. Higgins JPT, Cochrane C.*Cochrane handbook for systematic reviews of interventions* . 2019.

14. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: A systematic review and meta-analysis. *Soc Sci Med* . Nov 2017;192:125-133. doi:10.1016/j.socscimed.2017.09.021

15. Grimmett C, Corbett T, Brunet J, et al. Systematic review and meta-analysis of maintenance of physical activity behaviour change in cancer survivors. *Int J Behav Nutr Phys Act* . Apr 27 2019;16(1):37. doi:10.1186/s12966-019-0787-4

16. Corepal R, Tully MA, Kee F, Miller SJ, Hunter RF. Behavioural incentive interventions for health behaviour change in young people (5-18years old): A systematic review and meta-analysis. *Prev Med* . May 2018;110:55-66. doi:10.1016/j.ypmed.2018.02.004

17. Baskerville NB, Liddy C, Hogg W. Systematic review and meta-analysis of practice facilitation within primary care settings. *Annals of family medicine* . Jan-Feb 2012;10(1):63-74. doi:10.1370/afm.1312

18. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ (Clinical research ed)* . 2000;320(7247):1468-1468. doi:10.1136/bmj.320.7247.1468

19. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ* . 2019;364:k4817. doi:10.1136/bmj.k4817

20. Saramago P, Woods B, Weatherly H, et al. Methods for network meta-analysis of continuous outcomes using individual patient data: a case study in acupuncture for chronic pain.*Bmc Med Res Methodol* . 2016/10/06 2016;16(1):131. doi:10.1186/s12874-016-0224-1

21. Tuti T, Nzinga J, Njoroge M, et al. A systematic review of electronic audit and feedback: intervention effectiveness and use of behaviour change theory. *Implementation Science* . 2017/05/12 2017;12(1):61. doi:10.1186/s13012-017-0590-z

22. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Stat Med* . Sep 30 2013;32(22):3926-3943. doi:10.1002/sim.5831

23. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* . Jun 1992;48(2):577-85.

24. Petticrew M, Rehfuess E, Noyes J, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology* . 2013/11/01/ 2013;66(11):1230-1243. doi:https://doi.org/10.1016/j.jclinepi.2013.06.005

25. Davis J, Mengersen K, Bennett S, Mazerolle L. Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus* . 2014/09/10 2014;3(1):511. doi:10.1186/2193-1801-3-511

26. Tanner-Smith EE, Grant S. Meta-Analysis of Complex Interventions. *Annual Review of Public Health* . 2018;39(1):135-151. doi:10.1146/annurev-publhealth-040617-014112

27. Borenstein M, Higgins JP. Meta-analysis and subgroups. *Prevention science : the official journal of the Society for Prevention Research* . Apr 2013;14(2):134-43. doi:10.1007/s11121-013-0377-7

28. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted?*Statistics in medicine* . Jun 15 2002;21(11):1559-73. doi:10.1002/sim.1187

29. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?*BMJ* . 2017;356:j573. doi:10.1136/bmj.j573

30. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives.*Epidemiology* . Sep 1991;2(5):387-92.

31. Senn S. U is for unease: reasons for mistrusting overlap measures for reporting clinical trials.*Statistics in Biopharmaceutical Research* . 2011;3(2):302-309. doi:10.1198/sbr.2010.10024

32. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis.*Stat Med* . Nov 30 2000;19(22):3127-31. doi:10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.0.co;2-m

33. Sanchez-Meca J, Marin-Martinez F, Chacon-Moscoso S. Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological methods* . Dec 2003;8(4):448-67. doi:10.1037/1082-989x.8.4.448

34. Marin-Martinez F, Sanchez-Meca J. Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. *Educational and Psychological Measurement* . 2010/02/01 2009;70(1):56-73. doi:10.1177/0013164409344534

35. Laopaiboon M. Meta-analyses involving cluster randomization trials: a review of published literature in health care. *Stat Methods Med Res* . 2003;12(6):515-530. doi:10.1191/0962280203sm347oa

36. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Stat Med* . 2002;21(19):2971-2980. doi:10.1002/sim.1301

37. Walwyn R, Roberts C. Meta-analysis of standardised mean differences from randomised trials with treatment-related clustering associated with care providers.*Stat Med* . Mar 30 2017;36(7):1043-1067. doi:10.1002/sim.7186

38. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Statistical methods in medical research* . 2001/10/01 2001;10(5):325-338. doi:10.1177/096228020101000502

39. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. *Trials* . 2017/06/20 2017;18(3):280. doi:10.1186/s13063-017-1978-4

40. Miller JJ. The Inverse of the Freeman – Tukey Double Arcsine Transformation. *The American Statistician* . 1978/11/01 1978;32(4):138-138. doi:10.1080/00031305.1978.10479283

41. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data.*Stat Med* . 2010;29(29):3046-3067. doi:10.1002/sim.4040

Table 1: Units of randomisation and analysis for the 18 credible source comparison

	Number of studies	Number
Unit of randomisation Patient Health care professional Site (ward, hospital, surgery etc)	0 2 14	0 2 16
Unit of analysis Patient Health care professional Site (ward, hospital, surgery etc)	8 4 4	10 4 4

Table 2: Formulae to convert extracted data to SMDs

Type of outcome measure	Data to extract	Standardised mean difference (d)	Standard error of standardised mean difference
Continuous reported as mean or mean difference	Means (M_1 and M_2), standard deviations (S_1 and S_2) and sample size per group (n_1 and n_2)	$\frac{M_1 - M_2}{S}$ where S= $\sqrt{\frac{(n_1-1)^2 S_1+(n_2-1)^2 S_2}{n_1+n_1-2}}$=	$\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{d^2}{n_1+n_2-2}}$

Type of outcome measure	Data to extract	Standardised mean difference (d)	Standard error of standardised mean difference
Binary reported using odds ratios	Natural logarithm of odds ratio (lnOR) and standard error of log odds ratio SE_{lnOR}. This can be obtained from a 95% confidence interval for the odds ratio by taking natural logs and dividing by 2x1.96	$\frac{\sqrt{3}}{\pi}$lnOR	$\frac{\sqrt{3}}{\pi}SE_{lnOR}$
Raw binary data	Raw binary data (c_1/n_1 and c_2/n_2) where c_1 and c_2 are the number of participants complying with the behaviour of interest by group.	$OR = \frac{\frac{c_1}{n_1-c_1}}{\frac{c_2}{n_2-c_2}}$ Take natural log and continue as above	$SE_{lnOR} = \sqrt{\frac{1}{c_1} + \frac{1}{n_{1-}c_1} + \frac{1}{c_2} + \frac{1}{n_{2-}c_2}}$ Continue as above

Table 3: Summary of results for credible source data by 5 different methods

Method	**Number of comparisons**	**Result**	**Measure of heterogeneity**
Method 1 SMDs. Weights based on adjusted standard errors	18	Fixed effects SMD 0.14(95% CI 0.10 to 0.17) Random effects SMD 0.31(95% CI 0.14 to 0.51)	I^2 = 95.3%
Method 2 Separate analyses for binary and continuous data. Weights based on adjusted standard errors	OR 12 SMD 7	Fixed effects OR 1.13(95% CI 1.06 to 1.20) SMD 0.50(95% CI 0.42 to 0.59) Random effects OR 1.13(95% CI 1.06 to 1.20) SMD 0.92(95% CI 0.11 to 1.73)	I^2 = 0% I^2 = 98.0%
Method 3 SMSs. Weighting by number of HCP	18	SMD 0.57(0.50 to 0.64)	I^2 = 98.5%
Method 4 Albatross plot	18	All studies reported a positive effect so clear evidence of treatment effect.	Points not clustered around a single contour line so high levels of heterogeneity

Table 4 **Strengths and weaknesses of each approach**

Approach	Strengths	Weaknesses
Method 1 SMDs. Weights based on adjusted standard errors	All available data combined Clustering accounted for at level of randomisation	Mixture of different outcomes and formats likely to lead to heterogeneity May be difficult to interpret Inconsistent units of analysis (patient/HCP/site) Estimation assumptions may not hold
Method 2 Separate analyses for binary and continuous data. Weights based on adjusted standard errors	Likely to lead to less heterogeneity than method 1 as more similar measures are being combined. Little manipulation or estimation required	Does not combine all available information in a single analysis, which leads to loss of power and multiplicity Two analyses may give conflicting results
Method 3 SMDs. Weighting by number of HCP	Consistent units – weighted by health care professional	Number of health care professionals not always reported, requiring an estimate to be imputed Weighting may be related to quality of reporting; e.g. poorly reported studies get less weight. Unit of analysis error when not randomised at level of analysis Weights related to the size of the study but not the variability/precision Issues with SMDs as above
Method 4 Albatross plot	May include additional studies that report p-value only No assumptions	Difficult to check that p-values are correct if not accompanied by other summary data P-values prone to selective reporting Need to adjust sample size in some way for cluster trials

Figure Albatross plot using 'number of health care professionals' as sample size.