Article:
Turner, EL, Platt, AC, Gallis, JA et al. (82 more authors) (2021) Completeness of reporting and risks of overstating impact in cluster randomised trials: a systematic review. The Lancet Global Health, 9 (8). e1163-e1168. ISSN 2214-109X

https://doi.org/10.1016/s2214-109x(21)00200-x

Completeness of reporting and risks of overstating impact in cluster randomised trials: a systematic review

*Elizabeth L Turner, Alyssa C Platt, John A Gallis, Kaitlin Tetreault, Christina Easter, Joanne E McKenzie, Stephen Nash, Andrew B Forbes, Karla Hemming, on behalf of the CRT Binary Outcome Reporting Group**

Overstating the impact of interventions through incomplete or inaccurate reporting can lead to inappropriate scale-up of interventions with low impact. Accurate reporting of the impact of interventions is of great importance in global health research to protect scarce resources. In global health, the cluster randomised trial design is commonly used to evaluate complex, multicomponent interventions, and outcomes are often binary. Complete reporting of impact for binary outcomes means reporting both relative and absolute measures. We did a systematic review to assess reporting practices and potential to overstate impact in contemporary cluster randomised trials with binary primary outcome. We included all reports registered in the Cochrane Central Register of Controlled Trials of two-arm parallel cluster randomised trials with at least one binary primary outcome that were published in 2017. Of 73 cluster randomised trials, most (60 [82%]) showed incomplete reporting. Of 64 cluster randomised trials for which it was possible to evaluate, most (40 [63%]) reported results in such a way that impact could be overstated. Care is needed to report complete evidence of impact for the many interventions evaluated using the cluster randomised trial design worldwide.

Lancet Glob Health 2021; 9: e1163–68

*Members of the CRT Binary Outcome Reporting Group are listed at the end of the paper

Department of Biostatistics and Bioinformatics, and Duke Global Health Institute, Duke University, Durham, NC, USA (E L Turner PhD, A C Platt MA, J A Gallis ScM, K Tetreault MB); **Institute of Applied Health Research, University of Birmingham, Birmingham, UK** (C Easter MSc, Prof K Hemming PhD); **School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia** (J E McKenzie PhD, Prof A B Forbes PhD); **Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK** (S Nash MSc)

Correspondence to: Dr Elizabeth L Turner, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA **liz.turner@duke.edu**

Introduction

Well conducted research is important in all settings, but particularly so in global health, in which the accountability of researchers conducting "public health somewhere else" is ever more pressing.[1] Well conducted randomised trials are required to evaluate the community-based, multicomponent, and complex interventions that are so often of interest in global health research. Such interventions are typically evaluated using the cluster randomised trial (CRT) design.[2] Given the importance of these CRT evaluations for decision making regarding the adoption and scale-up of interventions, it is important that impact be correctly estimated by accounting for clustering of outcomes and that impact be clearly communicated. For the binary outcomes commonly used in CRTs,[3] absolute (eg, risk difference) and relative (eg, risk ratio) measures are appropriate and provide complementary evidence. The joint use of absolute and relative measures has been recommended by the Consolidated Standards of Reporting Trials (CONSORT) statement on reporting of results of randomised trials, and subsequently of CRTs, for nearly two decades.[4,5]

The rationale for recommending the reporting of both absolute and relative effects is well documented.[6–8] First, there is the potential for large relative effects to mask effects that have low public health relevance and, hence, for the intervention effect to be overstated. This situation arises when outcomes are rare (eg, <10% risk), and no accompanying absolute measure is presented.[6–8] For example, in a CRT of an intervention to promote tobacco cessation in India, the reported relative risk of 5·32 (95% CI 1·43–19·74) and absolute risk difference of 2·1 percentage points (95% CI 0·7–3·5) was obtained from a comparison of cessation proportions of 2·6% in the intervention arm versus 0·5% in the control arm.[9] Such a magnitude of absolute difference could be deemed to be of insufficient public health benefit for the intervention to be scaled up, yet had the relative effect been reported alone in this rare outcome setting, some users of the research findings might have overstated the intervention impact. Second, there is the potential for the effect of an intervention to be overstated when an odds ratio is used as a relative measure of effect. This situation arises when outcomes are common (eg, >10% risk) and the odds ratio is misinterpreted as a risk ratio.[10,11] For example, in a CRT of an intervention to link and retain adults in HIV care in Eswatini, the reported relative risk of 1·52 (95% CI 1·19–1·96) was obtained from a comparison of proportions of 64% in the intervention arm versus 43% in the control arm.[12] If the authors had reported an odds ratio in this common outcome setting, it would have been approximately 2·36. Had this figure been interpreted as a risk ratio—for example, stating that the intervention led to a 136% increase in the proportion of study participants linked and retained in HIV care—it would be a considerable overstatement of impact given that the reported risk ratio estimate showed a relative increase of approximately 52%.

It is not yet known how many CRTs adhere to guidelines to report both relative and absolute measures, and reporting might be worse than for individually randomised controlled trials given the additional challenges of accounting for clustering in analysis of CRTs. A 2020 review showed that only 9% of individually randomised controlled trials presented both absolute and relative measures.[13] Given the importance of CRTs with binary outcome for evaluations in real-world contexts,[3,14,15] we did a systematic review of design, analysis, and reporting practices of CRTs with binary primary outcome. Our goals were to determine the types of measures used, the completeness of reporting, and the potential for impact to be overstated. Importantly, our investigation was not limited by country, context, condition, or type of intervention, but instead includes all identified CRTs to provide a full picture of global practice.

Methods

See Online for appendix

In brief, we included all reports registered in the Cochrane Central Register of Controlled Trials of two-arm parallel CRTs with at least one binary primary outcome that were published between Jan 1 and Dec 31, 2017. See appendix p 8 for additional details and rationale regarding eligibility, including the choice of 2017 and the restriction to two-arm parallel designs and to full-scale CRTs (ie, no pilot or feasibility CRTs). Data extraction was done in duplicate, by random assignment of each paper to two of more than 80 participants at three different CRT-focused workshops, all of whom were experts in statistical methods, and most (66%) of whom had experience with CRTs (appendix p 9). Final data were agreed by each pair and included consultation with lead study authors (ELT or KH), as needed. This strategy was adopted to enable the research to be done rapidly to high quality but without dedicated funding.[16] For more details, see the study protocol (appendix pp 25–38) and related materials (appendix pp 2–7, 39–50). This systematic review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.[17] For the PRISMA checklist, see appendix pp 49–50.

We report descriptive characteristics of the included studies (eg, study domain, country, and intervention type) and the participants doing the data extraction. We summarise the proportion of studies reporting absolute and relative effects, as well as the proportion of studies that reported the primary outcome in such a way that there was potential for overstating the magnitude of the reported intervention impact. The absolute measures were: (1) difference measures (ie, prevalence difference, risk difference, or difference in proportions); (2) number needed to treat; and (3) other absolute measures. The relative measures were: (1) odds ratio; (2) risk ratio, relative risk, prevalence ratio, or ratio of proportions; and (3) other relative measures. Potential for overstatement of impact was defined as either of the following two conditions: the outcome was rare (≤10% risk in either trial arm) and only a relative measure was reported; or, the outcome was common (>10% risk in both arms) and the odds ratio was reported (see Introduction for examples of each of these situations). The magnitude of potential for overstatement of impact through the use of odds ratios for common outcomes was quantified via the ratio of the odds ratio and the risk ratio. This was estimated for each CRT by using the reported risks by arm to obtain both the odds ratio and risk ratio, from which the ratio of odds ratio to risk ratio was calculated. The overall goal of classifying each CRT according to whether there is the potential for overstatement is to draw attention to how the choice of measures could lead to overstatement of impact by users of the research findings. Relatedly, in this systematic review, we use the term impact to refer specifically to the reported effect estimates, and recognise that the true impact is a function of implementation, compliance, heterogeneity of effects, sustainability over time, and transportability to other contexts.

Results

In brief, after removing duplicates or reports published only on a trials registration website, 711 abstracts were screened, from which 89 articles were eligible for full-text review. Data were extracted by 82 individuals (appendix pp 9–10, 21–22) from the 73 articles determined eligible to be included the systematic review (appendix p 23; see appendix pp 51–54 for a full reference list). Comparisons of preworkshop independent data extraction showed that between-pair inter-rater agreement on up to 95 distinct variables exceeded 85% across the 82 individuals involved (appendix p 10), with a minimum of 84% at one workshop. Of the 73 CRTs in the systematic review, the most common conditions studied were infectious diseases (19 [26%]) and women's health (16 [22%]; appendix p 11). Close to half of the CRTs (35 [48%]) were done in at least one country designated as low income or middle income (appendix pp 11, 19), with the remainder having all study sites in countries designated as high income (38 [52%]).[18] Most CRTs selected a health facility or provider (41 [56%]) or geographical area (14 [19%]) as the cluster (appendix pp 11–12). Finally, most CRTs (46 [65%]) studied interventions targeted at participants, namely health promotion or educational interventions or direct participant therapeutic interventions (appendix pp 11–12). Follow-up data were typically collected using a questionnaire or survey (34 [47%]) or via electronic health records (22 [30%]; appendix p 13). Most CRTs (52 [71%]) enrolled fewer than 40 clusters and median cluster size was 48 (IQR 20–220; appendix pp 11–12). Nine (12%) of the 73 CRTs did not account for clustering in analysis (appendix p 15). 37 (77%) of the 48 journals in which the articles appeared endorsed the CONSORT statement on trial reporting (appendix pp 16–17) and 58 (79%) of the 73 CRTs appeared in one of those journals (appendix p 18).

Of the 73 studies included in the systematic review, few (13 [18%]) reported both a relative and absolute measure and so, overall, most (60 [82%]) provided an incomplete picture of evidence of intervention impact (figure 1; appendix p 14). Instead, within the main text of each article, most studies (46 [63%]) reported a relative measure only, eight (11%) reported an absolute measure only, and six (8%) reported no effect measure (typically only reporting proportions by arm), with a larger number of studies (15 [21%]) reporting no effect measure in the abstract. In addition to the six studies reporting no effect measure anywhere in the manuscript, three articles (4%) did not report outcome proportions by arm and therefore only 64 articles could be evaluated for the potential to overstate intervention impact (figure 1). Of the 59 CRTs (81%) reporting a relative measure, most

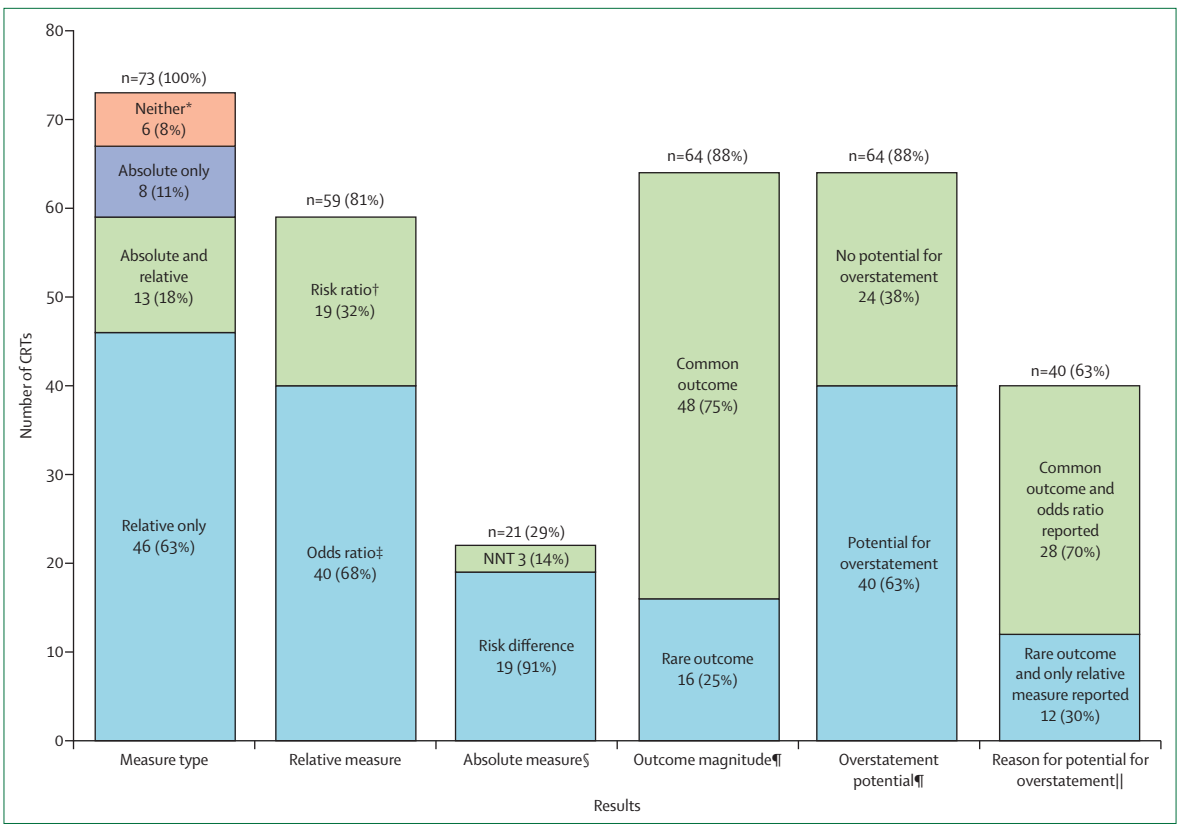

Figure 1: **Summary of reporting of results for CRTs with binary primary outcome**
A total of 73 CRTs were included in the systematic review. Percentages above each bar are out of 73, except for the final bar, which is out of 64. Percentages within bars have been rounded and therefore might not total 100% (whereas the percentages within the absolute measure bar do not total 100% due to overlap between the categories). CRT=cluster randomised trial. NNT=number needed to treat. ROR=ratio of odds ratios. *One article reported proportions per arm with a p value; four reported only proportions per arm with neither a p value nor other measure of statistical significance; one reported only a p value with no proportions by arm. †Two articles reported difference-in-differences as the between-arm (ie, intervention *vs* control) difference in the within-arm change in proportion from baseline to endline. ‡Two articles reported an ROR: one ROR was a ratio for intervention arm *vs* control arm of the within-arm odds ratio for baseline to endline change, and one ROR was the ratio between two levels of a postrandomisation covariate of the between-arm odds ratio (intervention *vs* control) of the intervention effect. §Categories are not mutually exclusive; one paper reports both an NNT and a risk difference. ¶For the 64 articles that report both an intervention effect as well as outcome proportions by arm. ||Among the 64 articles that report both an intervention effect as well as outcome proportions by arm.

(40 [68%]) reported an odds ratio with fewer (19 [32%]) reporting a risk ratio. Of the 21 studies (29%) that reported an absolute measure of effect, most (19 [90%]) reported a risk difference and three (14%) reported the number needed to treat (one of which also reported a risk difference)

Of the 64 CRTs reporting an effect measure with accompanying risks by arm, most (40 [63%]) were classified as having the potential for users of the research to overstate intervention impact (figure 1; appendix p 14). Potential overstatement was primarily (28 [70%] of 40 CRTs) because the odds ratio was the chosen relative measure for a common outcome (ie, with a reference risk >10%). In the remaining 12 CRTs (30%), potential overstatement was because only a relative measure (odds ratio, risk ratio, or similar) was reported for a rare outcome (ie, with a reference risk ≤10%). The magnitude of this potential for overstatement is illustrated for 55 of the 59 studies that reported a relative measure (figure 2) and was quantified via the ratio of the odds ratio and the risk ratio: the mean ratio of odds ratio to risk ratio was 1·4 (SD 0·6); median 1·21 (IQR 1·06–1·43, range 1–3·2). For the 28 CRTs in the common outcome setting that reported an odds ratio as the relative measure, the odds ratio averaged about 40% larger than the risk ratio and, in one case, reached three times the magnitude. Similarly, in the rare outcome setting with only a relative effect reported, those effects are typically of a large magnitude. For example, one CRT had a risk ratio of almost 25 and a reference risk of less than 5%.

Discussion

To prevent inefficient use of scarce resources, there is a duty to ensure that results of trials are reported in such a way as to allow for a clear and accurate understanding of the expected impact of an intervention. Our systematic review suggests that few CRTs with binary primary outcomes report both an absolute and relative measure of effect. As a consequence, most did not present evidence in a manner that facilitated accurate interpretation.[5]

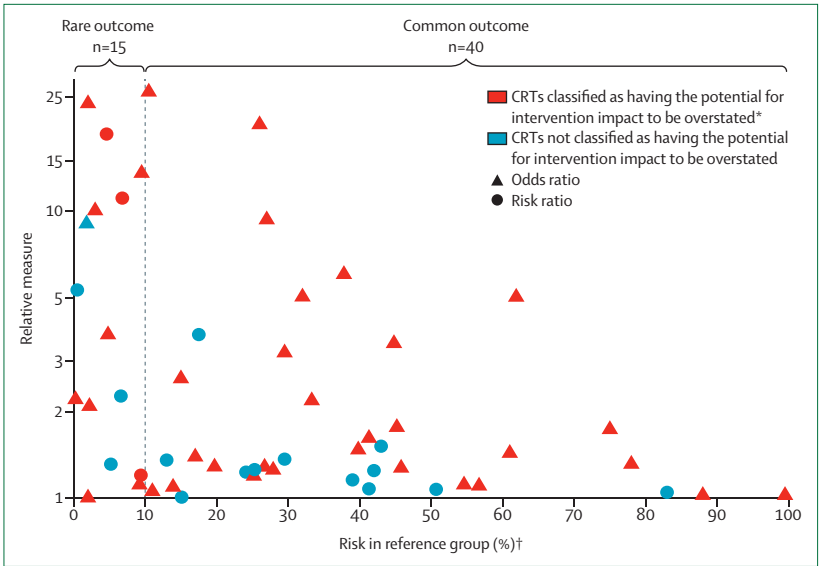

Figure 2: **Magnitude of relative effect in relation to reference risk for CRTs with binary primary outcome**
55 of the 59 articles reporting relative measures are included in this figure. Four articles cannot be included: two articles with a rare outcome (one that reported only a relative measure with an odds ratio of 512, and one that reported risk ratios, but only stratified by sex); and two articles with a common outcome which each reported an ROR—one ROR was a ratio for intervention arm *vs* control arm of the within-arm odds ratio for baseline to endline change, and one ROR was the ratio between two levels of a postrandomisation covariate of the between-arm odds ratio (intervention *vs* control) of the intervention effect. CRT=cluster randomised trial. ROR=ratio of odds ratios. *As shown in figure 1, 40 studies are classified as having the potential for intervention impact to be overstated: 12 studies reported a relative measure and had a rare outcome (reference risk ≤10%), 11 of which are shown here, and 28 studies reported an odds ratio and had a common outcome (reference risk >10%), all of which are shown here. †For ease of visualisation, the horizontal axis shows the reference risk, which is the smaller of the reported intervention-arm and control-arm risks.

Incomplete, and potentially distorted, presentation of the evidence has implications for policy and practice, possibly leading to the adoption of interventions with smaller than perceived impact. Given the widespread use of CRTs in the evaluation of interventions in a range of clinical and public health settings worldwide, such limitations could have far-reaching consequences.

The finding that relative effects are typically reported alone might be a consequence of the inherent challenges in analysing the clustered binary outcome data that often arise in CRTs.[11,19] Indeed, as mentioned above, we discovered that nine of the 73 CRTs did not account for clustering in analysis. Yet, methods are available to estimate both risk ratios and risk differences for clustered data. These methods include binomial mixed-effects models or generalised estimating equations, with log links (for risk ratios) or identity links (for risk differences), or cluster-level methods.[11] Relatedly, it might be surprising that relative effects were typically reported alone given that, as noted above, most of the journals in which the articles appeared endorsed the CONSORT statement on trial reporting, and most of the 73 CRTs appeared in one of those journals. As such, responsibility lies with journal editors to promote a shift in reporting practices. Likewise, there is a responsibility to avoid other poor practices identified in our systematic review, including: use of p values in baseline tables; ignoring clustering in sample size calculations; and lack of clarity about the primary outcome or the primary assessment timepoint (appendix p 13).

The finding that the odds ratio is the most commonly reported relative effect is probably explained by the fact that logistic regression mixed-effects modelling is the most well known approach to analysing clustered binary outcome data.[19] Although odds ratios can be difficult to interpret and can lead to overstating of impact, they have several advantageous properties. First, although the absolute impact of an intervention can differ by underlying risk, odds ratios are often stable across underlying risk and therefore might be more applicable in different populations or different contexts.[20] Moreover, for outcomes that are very common (eg, risk >80%), the odds ratio might be preferred to the risk ratio as a relative measure because the risk ratio can mask the magnitude of some effects; this is because absolute risk is bounded above by 100%.[21] Compared with their use in regular individually randomised trials, there are additional challenges to interpretation of evidence from odds ratios in CRTs, because intervention effects estimated by mixed models and generalised estimating equations using a logit link must be interpreted differently. That is to say, odds ratios estimated by use of mixed models should be interpreted as cluster-specific effects, whereas odds ratios estimated by use of generalised estimating equations should be interpreted as population-averaged effects.[2,15] The greater the variability between clusters, the greater the deviation of the cluster-specific odds ratio from the population-averaged odds ratio. By contrast, when identity links are used to estimate risk differences or log links are used to estimate risk ratios, the population-averaged and cluster-specific treatment effects are identical.[11]

Potential limitations of our systematic review include methodology and classification of overstatement. Regarding methodology, although we do not know how the relatively new crowd-sourcing approach would compare with a standard systematic review that relies on a few data extractors, the crowd-sourcing approach has been previously used by our team to evaluate quality of reporting of stepped-wedge designs.[16] To promote high-quality data reporting, data extraction (appendix pp 39–48) focused on mostly objective measures, data extraction was done in duplicate by individuals with statistical expertise (of whom most had CRT experience), and all pairs of data extractors were able to consult with one of two senior authors during in-person workshops for data finalisation. Nevertheless, because we leveraged expertise from participants at UK-based and US-based workshops, few authors would be considered experts in global health research done in resource-constrained settings. Regarding our classification of overstatement, we focused on an objective measure of potential overstatement rather than actual misinterpretation, which would be very difficult to

measure. As such, our estimate of 63% of CRTs with potential for overstatement of impact might be larger than what actually occurs in practice. In other considerations, although we limited ourselves to two-arm parallel CRT designs, our findings probably extend to more complex designs, including stepped-wedge, multiarm, and crossover CRTs, as well as to observational studies with binary outcomes. Similarly, although we have not examined the nature of reporting of effects for safety outcomes or of findings from subgroup analyses, similar attention should be paid to avoid overstating impact in these important domains.

With the increasing move to use evidence generated from CRTs to make decisions regarding the adoption of interventions in settings around the world, it is important that intervention impact be correctly and clearly communicated. For the binary outcomes so commonly used in decision making, reporting of both relative and absolute measures of effect is necessary to provide complementary and complete information. Many researchers do not realise the importance of this fact when communicating their study findings. Statisticians face some difficulties in estimating these effect measures in CRTs due to the complex nature of the models involved, but these difficulties are usually surmountable. Journals have a duty to ensure that published trials adhere to consensus-based reporting guidelines and to ensure that both relative and absolute measures are reported. If these issues are not rectified, there will undoubtedly be negative consequences on the evidence used for decision making about the adoption and scale-up of interventions around the world.

Contributors

ELT and KH led the development of the project and wrote the first draft of the paper. ACP developed the data extraction tools with input from JAG and ELT. All named authors (other than JEM) did the initial abstract screen. ELT, ACP, JAG, and KT processed the data and did the statistical analysis. ELT and ACP invited participants and emailed relevant material. All group authors and named authors (other than JEM and ABF) extracted data from a published paper reporting a CRT included in this systematic review, participated in an in-person data reconciliation workshop, and were invited to comment on the draft manuscript. All named authors approved the final manuscript. All authors had access to all data used in the study, and ELT, JAG, ACP, and KT accessed and verified all data.

CRT Binary Outcome Reporting Group

Affiliations for members are listed in the appendix (pp 55–56). *London, UK Workshop*: *Germany* C Adrion. *UK* N Akooji, A Bhangu, B Bridgwood, E Budgell, M Campbell, C L Chan, M Collinson, A Copas, S Eldridge, A S Forster, A Girling, J Glasbey, B Goulao, S Hackett, T Hamborg, R Hooper, K James, C I Jarvis, B Jones, B C Kahan, M Kanaan, L Kendall, C Kristunas, C Leyrat, S J Macneill, V W Madurasinghe, J Martin, L M Mwandigha, D Nepogodiev, O Omar, L A Pankhurst, H Perry, I Rombach, B Stuart, J A Thompson, A P Wagner, N Wilson. *France* A Caille, A M Mbekwe Yepnang, E Tavernier, B Giraudeau, L A Ndounga Diakou. *Canada* M Taljaard, S N Dixon. *Netherlands* M Moerbeek. *Australia* K L Grantham, J Kasza. *Durham, NC, USA Workshop*: *USA* S Cao, M Harding, K Kusibab, H-J Lee, K McCormack, K Moran, A Parish, R Simmons, T Truong, J R Vissoci, T Wang, X Wang, J Weber, J Wilson, S Yang, Z Yang. *Birmingham, UK Workshop*: *UK* H Bensoussane, J Bishop, V Cheed, A Gill, K Handley, P Hardy, C A Hewitt, N Ives, S Mehta, S Patel, Y Sun, R Woolley.

Declaration of interests

We declare no competing interests.

Data sharing

Data can be requested from the first author. The study protocol and statistical analysis plan are provided in the appendix (pp 25–28), together with a dictionary outlining the database and corresponding fields (pp 39–48).

Acknowledgments

We received no funding specifically for this systematic review. ELT is funded in part by awards R01-AI141444 from the National Institute of Allergy and Infectious Diseases and R01-MH120649 from the US National Institute of Mental Health; both Institutes are part of the National Institutes of Health (NIH). JAG and ACP's support of this project was made possible (in part) by grant number UL1TR002553 from the National Center for Advancing Translational Sciences of the NIH, and the NIH Roadmap for Medical Research. JEM is supported by an Australian National Health and Medical Research Council Career Development Fellowship (APP1143429). SN was supported by an award that is jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement, and part of the EDCTP2 programme supported by the European Union (grant reference MR/R010161/1). ABF acknowledges funding support from the National Health and Medical Research Council of Australia (grant ID 1183303). KH is funded by a National Institute for Health Research Senior Research Fellowship SRF-2017-10-002. The contents of the research included in this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of any of the funders. The research contributed by all authors of this manuscript are independent of their funders. Specifically, the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We wish to thank the three reviewers for their insightful comments and constructive feedback.

References

1 King NB, Koski A. Defining global health as public health somewhere else. *BMJ Glob Health* 2020; **5:** e002172.
2 Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester: John Wiley & Sons, 2012.
3 Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* 2016; **17:** 72.
4 Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340:** c332.
5 Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; **345:** e5661.
6 Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. *J Gen Intern Med* 1993; **8:** 543–48.
7 Barratt A, Wyer PC, Hatala R, et al. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ* 2004; **171:** 353–58.
8 Perneger TV, Agoritsas T. Doctors and patients' susceptibility to framing bias: a randomized trial. *J Gen Intern Med* 2011; **26:** 1411–17.
9 Sarkar BK, West R, Arora M, Ahluwalia JS, Reddy KS, Shahab L. Effectiveness of a brief community outreach tobacco cessation intervention in India: a cluster-randomised controlled trial (the BABEX Trial). *Thorax* 2017; **72:** 167–73.
10 Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *N Engl J Med* 1999; **341:** 279–83.
11 Gallis JA, Turner EL. Relative measures of association for binary outcomes: challenges and recommendations for the global health researcher. *Ann Glob Health* 2019; **85:** 137.
12 McNairy ML, Lamb MR, Gachuhi AB, et al. Effectiveness of a combination strategy for linkage and retention in adult HIV care in Swaziland: the Link4Health cluster randomized trial. *PLoS Med* 2017; **14:** e1002420.

13 Rombach I, Knight R, Peckham N, Stokes JR, Cook JA. Current practice in analysing and reporting binary outcome data—a review of randomised controlled trial reports. *BMC Med* 2020; **18:** 147.
14 Ivers NM, Taljaard M, Dixon S, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ* 2011; **343:** d5886.
15 Hayes RJ, Moulton LH. Cluster randomised trials. London: Chapman and Hall/CRC, 2017.
16 Hemming K, Carroll K, Thompson J, et al. Quality of stepped-wedge trial reporting can be reliably assessed using an updated CONSORT: crowd-sourcing systematic review. *J Clin Epidemiol* 2019; **107:** 77–88.
17 Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; **151:** 264–69.
18 Organisation for Economic Co-operation and Development. Classification of low and middle income countries. https://wellcome.org/grantfunding/guidance/low-and-middle-incomecountries (accessed March 18, 2021).
19 Li B, Lingsma HF, Steyerberg EW, Lesaffre E. Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. *BMC Med Res Methodol* 2011; **11:** 77.
20 Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2002; **21:** 1575–600.
21 Cook TD. Advanced statistics: up with odds ratios! A case for odds ratios when outcomes are common. *Acad Emerg Med* 2002; **9:** 1430–34.